|CURIEUX|

ACADEMIC REVIEW

October Issue

Issue 12 Part 3

Editing Staff

Chief-In-Editor

George Hasnah

Lead Editor

Amanuel Geremew

Ashley Du

Assisting Editors

Jing Yan Huang

Nidhi Potineni

Sophia Kong

Tony Yuan

Ridit Aggarwal

Manu Sharma

Rishi Gadikota

Eric Wang

Table Of Contents

Page 6: Quality of Dental Prosthetics and Daily Wellbeing: A Survey of Elderly Patients in the Chicagoland Area By Khushi Thakkar

Page 19: Social Comparison as a Mediator Between Social Media Use and Adolescent Mental Health By Leanne Mathews

Page 31: Jump or Drift? Comparative Empirical Analysis of Black-Scholes and Merton Jump-Diffusion Option Pricing Across Market Regimes By Sachin Srivatsav Sathyanarayanan

Page 44: Pool-Based Active Learning for Efficient AI-Assisted Lightweight CubeSat Structural Optimization By Akhil Tirumalaseti

Page 63: Applications of Artificial Intelligence in Eczema Management: A Scientific Literature Review By Aastha Chivakula

Page 83: The Nature and Nurture of Schizophrenia: A Review of Gene–Environment Interactions By Gurkriti Kochhar

Page 100: From Colonial Exploitation to Child Labor: The Enduring Cycle of Poverty in West Africa's Cocoa Industry By Veronica Wang

Page 106: Beyond Intuition: Exploring the Role of Probability and Statistics in Tennis By Neil Reddy

Page 119: The Use and Interpretation of Popular Memes (Geng) on the Chinese Short-Video Platform Douyin: A Qualitative Sociolinguistic Study of Chinese Short-Video Culture By Ki Kwong Kenrich Chen

Page 130: The Role of Environmental Stressors on the Increased Risk of Polycystic Ovary Syndrome By Niyanthri Jayakanthan

Page 137: How Effective Are Convolutional Neural Networks in Integrating Visual and Structured Data for House Price Prediction By Grace Sehgal

Page 146: Potential of Cell Towers By Vaibhav Gaddam

Page 151: A Mathematical Proof of the Necessity of Individualized Dosing Intervals in Patients with Reduced Metabolism By JuneYoung Park

Page 157: The Effect of Graphene Oxide Concentration on the Growth Dynamics and Apparent Amoxicillin Susceptibility of *Lactobacillus acidophilus* During Serial Subculture By Tengwen Lou and Xinyi Liu

Page 172: Domestic Stability, International Distrust: The FFKR's Post-2022 Brand Rehabilitation By Maria Runje

Page 187: Data Quality as A Foundation for Reliable Artificial Intelligence By Zara Vivekanand

Page 200: How Have FDA Breakthrough Therapy Designations for Psilocybin-assisted Therapy Between 2018–2024 Shifted Clinical Trial Design Standards and Psychiatrist Attitudes Toward Psychedelic Medicine in the US? By Srikar Kesanam

Page 218: Gene Editing and Synaptic Repair: Exploring CRISPR Applications in Restoring Neuronal Connectivity By Trishha Behal

Page 229: Compact Object Mergers as Laboratories for Extreme Matter, Curved Spacetime, and Fundamental Physics By Dhruv Sharma

Page 239: Smart Manufacturing and Industrial Automation: A Comparative Review of AI Adoption in Kazakhstan and the United States By Zhanibek Alikhan

Page 246: Effect of Microbial Necromass Deactivation Method on Radish Growth in Lunar Regolith Simulant By Shonik Shah

Page 269: Effects of Antioxidants in Reduction of Ethanol Induced Birth Defects By Vid Yash

Page 284: The Efficiency and Mission Feasibility of Nuclear Thermal Propulsion Compared with Chemical Propulsion for Crewed Missions to Mars By Alexandra Lovchinovskaya

Page 296: Utilizing Machine Learning to Test Novel Opioid Reversal Medication By Aayan Ghosh

Page 302: From "Being Seen" to "Being Recognized": Content Presentation and Audience Resonance among Visually Impaired Creators on Short-Video Platforms By Zitian Yang

Page 313: Predicting Bearing Failure 48 Hours in Advance: A Hybrid CNN-LSTM-Attention Approach on the IMS Dataset By Matt Wang

Page 330: Global Monetary and Central Bank Liquidity: Academic Foundations and the Current Global Landscape By Anya Patel, Mentor By Anamika Singh

Page 341: Circular Business Model Innovation and the Decoupling Question: A Case Study of H&M Group By Sravya Tatiraju

Page 356: In What Ways Can Kaitiakitanga Principles Provide Guidelines for Renewable Energy Projects in Te Tai Tokerau? By Aston Plunkett

Page 375: Physical and Psychological Challenges in Volleyball: Injury Prevention, Emotional Regulation, and Integrated Training Strategies By Andrew Huang

Page 381: The Influencer Impact: Social Media Influencers and Cosmetic Purchasing Behaviors and Habits Among Teenage Girls Ages 14–18 in Southern California By Grace Hirakawa

Page 395: A Coin Flip from Twelve Yards? The Penalty-Shootout "First-Kicker Advantage" at the World Cup, and Where the Pressure Really Bites By Robert Sultan

Page 406: The Imagined Gaze: Digital Beauty Standards, Celebrity Culture, and Eating Disorders Among Adolescent Girls By Martina Chai

Page 417: Extending Lithium-Ion Battery Lifetime: A Review of Failure Prediction and Degradation Mechanisms By Arav Choudhary, Advisor By Sanjeev Datta

Page 440: Modeling Exploratory GenAI Exposure Across the Animation Production Pipeline: A Scenario-Based Analysis By Santosh Dasari

Page 455: How Does Changing the Axis Ratio of a Logarithmic Potential System Affect Its Integrability? By Muzhi Fang

Page 470: The Impact of Limited English Proficiency on Healthcare Access Across Six Domains in the United States By Shruthi Shanker

Page 485: Research Shows A Strong Inverse Correlation Between Religiosity and Per-capita Spending on Education. Does One Cause the Other? By Seoyeon Lee

Page 491: Impact of Pitch Clock Implementation on Major League Baseball Engagement and Player Performance By Sajan Komanduri and Abdulla Kerimov PhD

Page 501: Potential of Piezoelectric Elements on Renewable Electrical Energy Generation By Benjamin Jetty and Henry Love

Page 512: Safety and Efficacy of Topical Acne Ingredients: A Narrative Review By Diya Rishi

Page 523: The Effects of Camber and Thickness on Aerodynamic Performance of NACA Four-Digit Airfoils: A Simulation Study with Experimental Flow Visualization By Landry Silvers

Page 540: Risk Factors For Skin Cancer By Navya Sampathkumar, Mentor By Dallin Lowder

Page 552: Factors Explaining Differences Between Club and International Performances By Abhinav Swaika

Page 560: Financial Sustainability, Regulation, and Competitive Balance in European Football Leagues By Hashem El Barbir

Page 570: Beyond Representation: Diversity Mandates Will Not Solve Cancer Inequality By Meghna Saha

Page 582: Voter Turnout: How Gen Z is Impacting Climate Policy By Carla Mansilla

Page 590: Breathing Breaks During Lecture-Style Classes: Impacts on High School Students' Short-Term Comprehension of New Astronomy Content By Shreya Bollapally

Page 598: The Capital Gains Tax: Its Impact on Socioeconomic Disparity in 21st Century America By Christopher Page

Page 612: The Impact of Song Lyrics on Emotional Response: A Multidisciplinary Perspective By Qianyi Wang

Page 625: From Craft to Content: How Narrative Shapes Preference in Fashion Consumption By Shanzay Rahman

Page 639: Comparing Exoplanet Transit Detection Pipelines Under Realistic Noise and Data Gaps By Kyle He

Page 651: Soft-Core Radicalization: Feminism and the Digital 'Trad-Wife' Aesthetic in Emergent Conservative Gen-Z Politics By Sainzaya Buyankhuu

Page 666: How do Marine Organisms Inspire Sustainable Product Design? By Aarohi Godge

Page 683: The Physics of Color Rendering and Emotional Impact: How LED Spectrum Optimization Transforms Audience Perception in Musical Theater By Hridaan Jhunjhunwala

Page 699: Self-Perception, Well-being, and Collaborative Decision-Making in Middle School Students: A Cross-Sectional Study by Grade and Gender By Stella Kakkanatt

Quality of Dental Prosthetics and Daily Wellbeing: A Survey of Elderly Patients in the Chicagoland Area By Khushi Thakkar

Abstract

Tooth loss is often treated as a normal part of aging, yet the everyday experience of the older adults who rely on dental prosthetics is frequently overlooked in favor of purely clinical measures. This study examined how the type and quality of dental prosthetics relate to the daily wellbeing of elderly patients living in nursing homes and assisted-living facilities in the Chicagoland area. Using a descriptive quantitative design, geriatric residents completed an eight-question survey addressing prosthetic type, comfort while eating and speaking, pain and irritation, social confidence, maintenance habits, and overall satisfaction. Responses were analyzed in Jamovi using descriptive statistics, Pearson correlations, and independent-samples t-tests. Comfort and pain levels varied notably across prosthetic types, and a positive relationship emerged between physical comfort and social confidence. Maintenance frequency showed a moderate correlation with satisfaction, and participants who had used their prosthetics longer tended to report better adaptation. By centering the real-world experiences of elderly patients rather than clinical outcomes alone, this study highlights the importance of prosthetic fit, accessible routine care, and patient education in supporting quality of life.

Introduction

For many older adults, tooth loss has become a normalized part of aging. Many of these older adults rely on dental prosthetics like dentures, partials, or implants to perform their daily functions despite losing teeth. Even though these devices were designed to make daily functions like eating, speaking and socializing easier, many prosthetic users still experience discomfort, difficulty chewing, or self-consciousness when wearing them. These challenges are even more noticeable in residents of nursing homes and assisted living facilities because they are constantly surrounded by their peers and their access to dental care tends to be more limited than for prosthetic users who live independently.

Although dental prosthetics are an essential component of the lives of older adults, many studies focus primarily on clinical measurements rather than patient experience. Most research studies focus mainly on clinical measurements like chewing efficiency, bone support or the success rate of prosthetics in a controlled setting instead of a patient's personal experience that includes all the components such as comfort and social/emotional experience, not just the technical components. These findings are extremely important to contributions/innovations to the dental field but they don't fully explain what daily life feels like for seniors who wear these prosthetics. This suggests a clear gap between what professional dentists/researchers consider a "successful" execution of treatment in terms of dental prosthetics and how seniors themselves experience comfort, confidence and ability to eat/chew and speak in their day to day life.

Because of this gap, my study aims to examine how the specific type and quality of dental prosthetic that an elderly patient wears connects to their chewing/speaking ability, comfort and confidence in a real life setting. By gathering self reported data from older adults in the Chicagoland area, this study seeks to contribute to an underrated but crucial understanding of how dental prosthetics shape daily life and where improvements can be made.

Literature Review

Theme 1: The Historical Development of Prosthetics and Its Influence on Modern Outcomes for Patients

A large amount of research showed that the expectations that senior patients have today about comfort, quality and chewing ability didn't appear out of nowhere. They were shaped by a history of trial and error when designing prosthetics. Abraham explained that humans have been trying to identify ways to replace missing teeth for centuries before modern dentistry evolved. Before modern titanium implants came into use, ancient homo sapiens would use materials like carved stone, seashells and early metals to replace the aesthetic and function of missing teeth. Although these early designs weren't perfect and took a great deal of effort to make, they set a precedent for the idea that dental prosthetics are not only meant to fill a space, but they also must restore function. A study conducted by Neves et al. adds to this idea through its analysis of historic dentures that were made of ivory and rubber. These early prosthetics were functional but extremely fragile, bulky and uncomfortable to wear. The conditions of these older dentures show how the experiences for patients have evolved a lot from the past and aren't as limited as they were before. Today, the standard for dental prosthetics that patients look for is much higher. Patients today look for a prosthetic that looks almost identical to a real tooth/set of teeth. They also want something that feels natural, allows them to eat comfortably; these factors combined are intended to boost the patients confidence. All of these historical studies together provide an explanation of how standards for comfort, aesthetics, function and durability are important evolving themes in prosthetic dentistry. They emphasized that the evolution of materials and care techniques used in prosthodontics has heightened the expectations of patients. This theme directly connects to the aim of this study which was to identify to what extent certain factors contribute to oral health outcomes in prosthetic users.

Theme 2: Modern Dental Technology and Shaping Expectations

With advancements being made in dental technology, scientists and prosthodontics changed their concentration on how changing the design of the prosthetic affects the comfort of the patients. Bors et al. experimented with the "smart" design of the prosthetics by using computerized scans and 3D prints. This helped them design their prosthetic more precisely which then led them to require less adjustments of the dentures and implants. The researchers did state that although these modern methods provided promising results, they were often costly and are only limited to large cities and advanced clinics. Another technical consideration that emerged

from this idea is the extent to which these types of prosthetics were able to match with the patients occlusion: Pokorny et al. studied the importance and relevance of studying traditional occlusion and to what extent it applies to modern implants. They discovered that there is a direct correlation between minor misalignments in bite and the discomfort that patients experienced while eating. Their findings suggested that comfort does partially depend on the material and shape of the dental prosthetic but it also hinges on how well these technical aspects support jaw movement and how that ultimately affects the comfort and function of a dental prosthetic. Overall, both of these studies suggested an emphasis on looking at modern technology and its potential to improve the daily life of senior patients, but only if the prosthetics were deliberately designed and fitted properly. These findings support the theme of my study, which is emphasizing the day to day importance of topics like precision and quality which are extremely important because they relate directly to older patients wearing their prosthetics daily.

Theme 3: Access Barriers That Contribute to Disparities in Dental Care

While I was analyzing prior research about the field of dental prosthetics and their outcomes on seniors' experiences, I observed that there is a consistent pattern among multiple studies: Glenza reported that many elderly patients in the United States skipped important dental appointments due to high out-of -pocket costs. Even basic maintenance visits became difficult for individuals who had budgets that they couldn't cross. Nasseh et al. also disproved the idea that insurance coverage always means actual access. The results of their study suggested that Medicare Advantage dental plans include many restrictions on usage. As a result, despite having health insurance on paper, many elderly patients who needed dentures or implants found it nearly impossible to pay for them due to their insurance plans. According to Zhao and Kim, extending dental coverage and attempting to reduce these insurance restrictions only addressed a part of the problems associated with access to dental care and prosthetics. There are factors such as lack of transportation, low number of clinics, and lack of knowledge/education about dental insurance requirements which also contributed to trouble accessing dental care for certain elderly populations which ultimately made it difficult for patients to obtain the dental prosthetics they needed. These studies suggest that both the quality of dental care and patients comfort depend not only on clinical considerations, but also on socioeconomic factors. In the case of elderly patients living in the Chicagoland area, this may especially be relevant because these factors can contribute to an access barrier to dental care and overall satisfaction with their prostheses. This strongly supports the need for research that provides insight on everyday patient experiences.

Theme 4: The Importance of Patient Oral Health Knowledge for Successful Outcomes

Another common theme I found across the literature is that seniors' understanding of their prosthetic and their oral health in general largely influences their oral health and satisfaction outcomes as much as the actual prosthetic itself. According to research by Atagün and Kalyoncuoğlu, while many seniors had heard of implants, most of them had major misconceptions about cost, procedure length and risk/side effects associated with them. These

misconceptions led to many of these patients to feel anxious and this ultimately caused them to refrain from investigating more about implants, even though these implants and other prosthetics could offer life changing opportunities to some patients.

Similarly, Yu et al. pointed out that oral health literacy directly correlates to how well older adults can maintain their dentures or implants with care. Patients who received more guidance were able to maintain their oral hygiene while using prosthetics and showed fewer problems with their dentures and implants than patients who did not receive adequate instructions. Those with lower oral health literacy reported to be less comfortable in their prosthetic and they even felt less confident than their more orally literate counterparts.

These results suggest that oral education and patient-provider communication are important factors when it comes to enhancing older adults' experiences with dental prostheses. This idea made up a large basis of my thinking that the subjective perceptions about prosthetics needed to be more prioritized than the objective/material outcomes specifically in this study.

Connection to the Present Study

Across all of these themes, the previous literature made one idea very clear: the quality/design of the prosthetics and access to dental care/patient understanding both heavily influence the outcomes when older adults use dental prosthetics. However, most of these previous studies were conducted in either healthcare facilities like dental offices or they focused more on technical outcomes like aesthetics instead of day-to-day experiences. This suggests that there is relatively little research that focuses on the quality of life among older adults within a specific demographic or geographic region such as the Chicagoland area. To overcome this gap, the study focused on gathering primary, self reported data from the senior population that lives in nursing homes and assisted living facilities. This allowed for a more in depth perspective about how the type and quality of prosthetic treatment relate to comfort, chewing ability, speaking ability, confidence and consistency of maintenance in a real world setting.

Method

The aim of this study was to examine the impact of the quality and type of dental prosthesis worn by seniors in the Chicagoland area on their quality of life. I chose to use a descriptive quantitative design to collect my data because this type of method allows me to analyze and describe a current situation without having to change or manipulate any variables. I used this research method by analyzing the characteristics of different patients' prosthetics and how this affected their day-to-day function. In this case, it helped me look for correlations between the type of prosthetic a patient wears and how the patient feels about their comfort, speaking ability, eating ability and confidence. This type of survey also allowed me to collect measurable data on real experiences rather than clinical based control settings. I based this decision on research done by Prakash et al. who surveyed over 200 elderly patients on their usage of dentures and implants to compare their comfort and confidence. Their study supports the idea that even a simple survey can show extremely important patterns that can help

professionals in the field of prosthetic dentistry. Similarly, Samiou et al. conducted cross sectional research in nursing homes to test how different types of prosthetics (like dentures, implants, bridges and crowns) specifically influenced chewing ability. In support of these types of analyses, Yu et al. conducted a meta-analysis, reviewing dozens of oral health literacy research papers; they found that questionnaires that measure perception are one of the most accurate ways to measure the habits and satisfaction of patients. These sources all led me to analyze self reported data instead of clinical/experimental measurements.

Context

The study was conducted in several nursing homes and assisted living facilities in the Chicagoland area. There are many geriatric patients in all of these facilities who solely rely on dental prosthetics like dentures or implants for their comfort and nutrition. These facilities are the ideal setting to collect research because unlike clinical environments, the setting is not as controlled and shows a wide variety of patients and prosthetics. In order for additional guidance, I collaborated with a licensed dentist who visits local senior facilities as their job daily. This partnership is extremely similar to the research team of Samiou et al., who also worked directly with nursing home dentists to ensure safe and professional communication with participants. Before beginning my data collection process, my entire research plan and questionnaire was reviewed and approved by an Institutional Review Board (see Appendix D) to confirm that the study was ethical. An IRB was extremely necessary for this study because human subjects are involved.

Participants

The participants in this study were chosen through purposive sampling, this means that I intentionally chose people who represented my topic instead of randomly selecting patients. This approach made the most sense for my study because I needed a very specific demographic: older adults who use dental prosthetics in their everyday lives. All of the participants were 65 years or older and resided in nursing homes or assisted living facilities in the Chicagoland area.

I chose this group because, according to Prakash et al., focusing on older patients who already use dentures or implants produces the most optimal and relevant data about their daily comfort. Focusing on a specific group also made this study more feasible due to the fact that the data can all be collected at one point; this reduces the possibility of my data collection process interfering with the daily lives of the participants.

Before beginning any data collection, each patient was given an informed consent form written in clear, straightforward language (See Appendix A). Informed consent forms were also used in the data collection process by Yu et al., who emphasized that surveys for older patients need to use easy-to- read language to prevent any misinterpretation. I read the questions out loud and recorded answers with permission for participants who had difficulty in reading or writing. Some residents also received assistance from nursing staff if they needed translation or mobility support.

I aimed for a sample size of 50-75 participants. This sample size was similar to the group sizes used by Samiou et al. in their study which was also conducted in nursing homes. They found that this is a balanced number of participants because it provides a wide variety of data/experiences to analyze but it also doesn't make the data collection process overwhelming. There were no personal indicators like names or room numbers, each participant's response was labeled with a number to protect confidentiality and follow HIPAA protocols.

Data Collection Materials

The primary data collection tool used for this study was a questionnaire (see Appendix B) that I developed after analyzing existing research about dental prosthetics and oral health related quality of life. I made sure my questionnaire included similar concepts to those previously tested for reliability. Prakash et al. used a ten-question survey to look at the comfort, chewing ability and confidence of patients who use dentures and implants.

My questionnaire included 3 sections (8 questions total):

Section A: Demographic/contextual information (type of prosthetic used)
Section B: Likert scale questions (1=*strongly disagree* to 5=*strongly agree*) about comfort, chewing and speaking ability and confidence.
Section C: A multiple-choice question regarding consistency with dental care and maintenance of prosthetics.

This questionnaire was designed in this way because Samiou et al. found that Likert scales allow for easier comparison across prosthetic types and levels of patient satisfaction. These measurement techniques directly address the research question by analyzing specific areas such as comfort, chewing, speaking ability, maintenance habits, and confidence.

Data Collection Procedures

After receiving IRB approval and permission from each care facility, I scheduled short sessions with the Dentist I collaborated with to conduct surveys during times that wouldn't interfere with the daily routines of patients. Before asking questions, I introduced myself, explained the purpose of the study and made sure that patients understood that their responses would be anonymous. I chose to conduct the survey in person to ensure that the data was consistent and reliable; each participant received the same questions in the same order from the same person (me). Surveys took about 6-10 minutes to complete depending on communication barriers.

For participants who were unable to read/write, I read each question out loud and recorded their answers.This approach was supported by a study done by Yu et al., who found that seniors tend to express themselves more thoughtfully when spoken to directly. This was also necessary for participants who had impairments that prevented them from reading and writing.

Surveys were conducted in quiet areas of the facilities such as activity spaces or libraries to ensure participant privacy and comfort, similar to how Taue et al. conducted their analysis of dental implants in private areas to reduce potential self-consciousness/embarrassment. These steps followed the ethical research/outreach procedures that are emphasized by the UIC College of Dentistry that helped me understand the importance of respect and confidentiality for elderly patients.

After collecting all responses, I double checked that each form was filled completely. All physical responses were stored in a secure folder. Next, I took the data from the folder and entered it into a password protected google sheet spreadsheet. I implemented strong security protocols: I was the only one with access to both the spreadsheet and the physical copies of the responses. These procedures were implemented to maintain ethical standards including HIPAA that mandates protection of patient privacy.The procedure was also standardized, as each participant completed the same survey, presented in the same order and the questions were asked by the same and only researcher. Both the survey and my informed consent form are included in Appendix A and Appendix B (respectively). This was done to ensure that future researchers can fully replicate this study.

Data Analysis

Once all responses were organized, I analyzed the data using Jamovi, a statistical software that helps create graphs and run tests like correlation and t-tests. I chose Jamovi because it allows for efficient analysis under the R programming model while still producing accurate results in a user friendly manner.

First, I calculated the standard deviations and means for each question to summarize overall trends and variation within the data.
Then, I ran conducted statistical tests:

1. The Pearson correlation test that measures the connection between two variables. I used this test to examine if higher prosthetic quality and quality of care was related to comfort, chewing ability and confidence. Samiou et al. also used Pearson correlation to examine relationships between types of dental prosthetics and chewing confidence.
2. The Independent Samples T-Test which is a test that compares the average values between two separate variables. I used this test to identify the differences in quality of life outcomes between denture users and implant users. Taue et al. also used this method in their study to compare oral function outcomes.

These statistical tests allowed me to identify relationships and differences between prosthetic type and dental care to their quality of life outcomes which directly addressed the research question.

Results and Findings

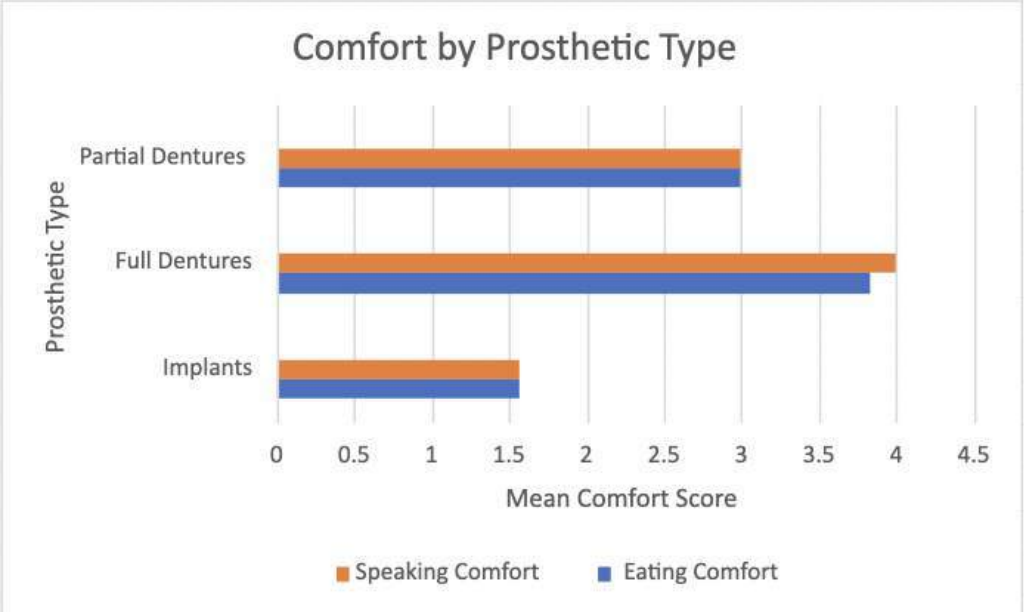
Overview of Data Analysis

This study investigated the relationship between dental prosthetics and the daily lived experiences of elderly nursing home/assisted living residents in the Chicagoland area. My quantitative survey data was analyzed using descriptive statistics and Pearson correlation analysis. My sample population was 70 patients aged above 65 who resided in various nursing homes in the Chicagoland area. The purpose of this study was to identify important patterns among the variables without manipulating any clinical or patent related factors. The statistical tests were carried out using a software called Jamovi. The full statistical output is presented in Appendix C. The results are presented by the main outcome categories to address the research question.

Comfort While Eating and Speaking Based on Prosthetic Types

The results of the comfort level with eating and speaking showed significant variations among the prosthetic types (see Figure 1). The patients who wore full dentures experienced the highest comfort levels while eating (M=3.83) and speaking (M=4.00), compared to the patients who wore partial dentures (M=3.00). On the other hand, the patients who wore dental implants showed the lowest comfort levels for eating (M=1.57). This pattern was supported by independent group comparisons: full denture users reported higher comfort and confidence levels than implant users, and partial denture users reported higher comfort than implant users as well.

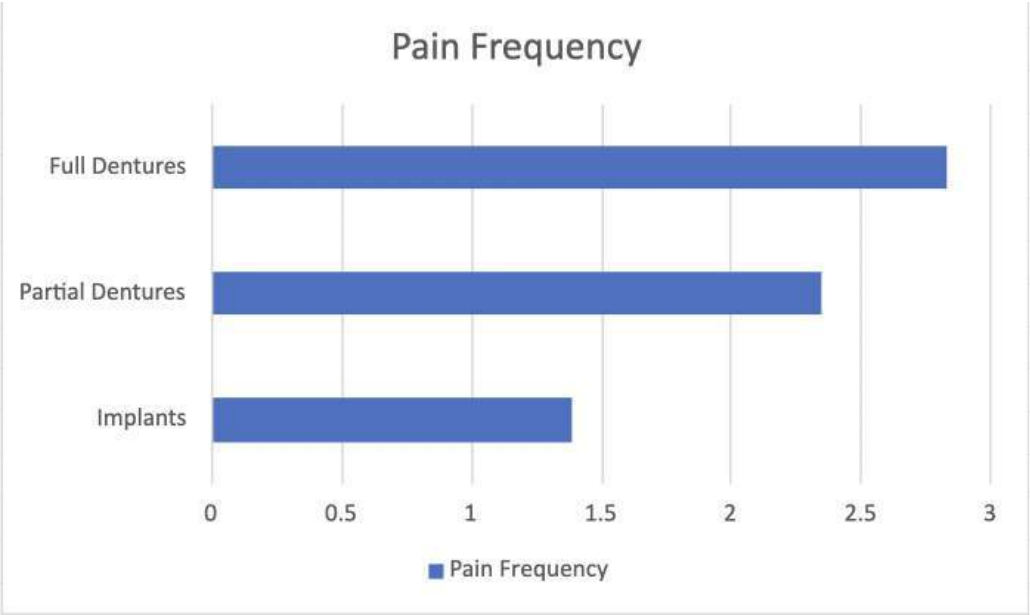
Figure 1. Comfort Levels by Prosthetic Type



Pain and Irritation Across Prosthetic Type

Pain frequency also showed great variation across prosthetic groups (see Figure 2). Implant users reported the highest average pain score (M=1.39) compared to the mean pain scores of partial denture users (M=2.35) and full denture users (M=2.83).

Figure 2. Pain Frequency by Prosthetic Type



Confidence and Social Comfort

Pearson correlation analysis showed a positive relationship between comfort and confidence. Participants who reported higher levels of comfort while eating and speaking were more likely to report higher levels of confidence in social settings (r =.925 and r=.912) whereas participants who experienced frequent discomfort and pain were also more likely to report no change in their confidence and some even reported decreased confidence.

Maintenance and Overall Satisfaction

Maintenance showed a more moderate pattern. Participants who attended regular dental visits for prosthetic maintenance were more likely to report slightly higher satisfaction with their prosthetic (r =.667) compared to those who sought care only when a significant problem occurred.

Time Since Treatment and Adaptation

My data also showed a pattern of time and prosthetic placement. Participants who had been using their prosthetic for longer periods of time reported greater ease with speaking (r =.930).

Interpretation of Key Findings

The purpose of this study was to examine how the type and quality of dental prosthetics directly influence the daily comfort, wellbeing and social confidence of patients over 65 residing in nursing homes and assisted living facilities in the Chicagoland area. The results of this study demonstrate a clear relationship between aspects of a patient's everyday experience when using prosthetics and multiple factors of quality of life within a real life, non-clinical setting that is often overlooked in existing research.

One of the most significant findings was the difference in comfort across prosthetic types. Participants who used full dentures reported the highest levels of comfort while eating (M=3.83) and speaking (M=4.00); contrary to implant users who reported the lowest levels of comfort in both categories (M=1.57). This statistic suggests that the type of prosthetic a patient uses (out of the ones being accounted for in this study: full dentures, partial dentures, and implants) plays an important role in how comfortable patients feel performing basic daily functions like eating and speaking. Eating and speaking are two essential functions/activities for older adults who rely on prosthetics because they are essential to maintaining independence and nutrition. This finding is important because although implants are usually seen as more technologically advanced and superior because they are permanently anchored to the jawbone, they tend to require more maintenance than partial or full dentures. This is particularly important because Nasseh et al. demonstrated that older adults often face access barriers to accessing consistent dental care despite having health insurance. This suggests an explanation for why implants were deemed as more uncomfortable in this study: elderly patients may not receive sufficient follow up dental appointments for adjustments and maintenance which can result in increased discomfort, particularly for implants that require ongoing care and precise adjustments.

Another factor that followed a similar pattern to eating and speaking ability was pain frequency. Implant users reported the highest levels of pain and irritation (M=1.39), compared to partial denture users (M=2.35) and full denture users (M=2.83). Although implants are seen as one of the most advanced forms of dental prosthetics, these results suggest that implant users in this particular sample experienced the highest level of discomfort. This difference may be attributed to factors such as healing time, prosthetic fit or access to follow up care. Because the study focused on nursing home residents, some patients often have limited access to frequent follow up adjustments that mitigate irritation over time. Building on the previously discussed idea that patients in nursing homes may face access barriers to consistent dental care, reducing the likelihood of regular follow-up appointments and necessary adjustments. This may contribute to higher levels of irritation and discomfort particularly for implants, which require ongoing maintenance.

Another significant finding was the relationship between physical comfort and social confidence. The Pearson correlation analysis revealed that eating comfort and confidence had a strong positive correlation (r=0.925), as did speaking comfort and confidence (r=0.912). Participants who reported higher comfort while eating and speaking were more likely to have higher confidence in social settings. This relationship is particularly meaningful because social interaction is a central component of daily life in nursing homes and assisted living settings. If

patients do not feel comfortable eating or speaking in public, then they might be less inclined to eat or speak in front of their peers. This may lead to avoidance of social interaction and reduced participation in group activities which can reduce social confidence and overall wellbeing.
My data also revealed that there was a moderate correlation between prosthetic maintenance and satisfaction (r=0.667). Those participants that had more frequent dental visits for maintenance were more likely to have higher prosthetic satisfaction than those who only visited their dentist when there was a problem with their prosthetic device. This aligns with research by Prakash et al. which suggests that a part of overall quality of life in the elderly hinges on oral functions like comfort while chewing and speaking. This suggests that regular maintenance visits may improve prosthetic fit and function which may contribute to higher levels of satisfaction. These findings highlight the importance of proactive maintenance strategies such as establishing systems of routine and accessible dental care, particularly for patients in care facilities to improve their overall long term comfort and satisfaction with their prosthetic.

Another pattern emerged when analyzing the length of time participants had used their prosthetic. My data showed a strong relationship between time since treatment and speaking comfort (r=.930). Participants who had used their prosthetic for longer periods reported greater ease in speaking. One possible explanation for this is that individuals may have gradually adapted to their prosthetic as they became more familiar with it during daily activities. This suggests that initial discomfort may decrease over time and that early struggles may not entirely reflect long-term outcomes. These findings highlight the importance of an adjustment period and early support for prosthetic users in order to better manage patient expectations and provide overall satisfaction.

Connection to Previous Research

While all these studies identified important outcomes, they primarily focused on clinical performance. The current study extends this work by examining personal experiences and effects on patients who use dental prosthetics. Factors including confidence, comfort and social engagement also hinge on the quality of dental prosthetics. By focusing on self-reported experiences in real world settings, this study addresses a gap in existing literature that often overlooks the lived experiences of elderly prosthetic users who may face limitations in access to adequate dental care.

Limitations

Several limitations should be considered when interpreting my findings. For starters, the data was self reported, which may introduce bias, as participants may overestimate or underestimate their comfort levels or misinterpret survey questions. Another limitation is that my study didn't have any control for factors such as age, access to follow up care and oral health history which may have influenced the relationships observed between prosthetic type and quality of life measures. Another limitation is the study sample: participants were drawn from nursing homes and assisted living facilities in the Chicagoland area. While this population

provided useful insight into the experiences of these prosthetic users, the findings may not represent elderly individuals living independently or in other areas as individuals living independently may have greater access to dental care and different daily experiences with prosthetic use. Further studies that include a larger and more diverse sample could help determine if similar patterns appeared in different populations and settings. Despite these limitations, there were various steps taken to ensure consistency, including the use of a standardized questionnaire, conducting all the surveys by the same researcher and obtaining IRB approval and following all HIPAA protocols to maintain an ethical data collection process.

Conclusion

This study demonstrates that the type and quality of dental prosthetics are closely related to how elderly patients live their daily life in nursing homes and assisted living facilities. The results suggest that prosthetic type has a heavy influence on day to day comfort, as patients with full dentures reported higher levels of eating and speaking comfort compared to those with implants. Although this study does not establish causation, it does suggest that clinical definitions of successful prosthetics do not always reflect how patients actually feel using them in everyday situations.

The results also showed that a strong positive relationship exists between comfort and confidence. Comfort levels in both eating and speaking were associated with higher levels of confidence in social settings. In nursing homes, it is common for patients to share common facilities like dining rooms and lounges. Discomfort while eating and speaking may affect patients' desire to be involved in social interactions with their peers. Therefore, this suggests that the quality of an individual's dental prosthetic has both physical and psychological outcomes. Maintenance habits were also identified as a variable that correlates with patient satisfaction. Patients who maintained regular visits for their prosthetic adjustments were more likely to show higher levels of overall satisfaction with their prosthetic compared to those patients who only visited a dental provider when they experienced issues with their prosthetic. In addition, patients who utilized their prosthetics for a longer period of time showed higher levels of speaking comfort, indicating that patients may be able to adapt to their prosthetics over time.
By taking an approach that emphasizes patient experiences in real life settings, this study provides further information beyond clinical outcomes of how dental prosthetics impact the lives of older adults. Further researchers could expand on this study by examining a larger sample size/group of elderly prosthetic users. Further studies could also compare patient self reported comfort levels with clinical measures of prosthetic function. Such studies could help provide a more complete understanding of how dental prosthetics affect the overall oral health in the elderly populations by examining how prosthetic materials, cost barriers and access to dental care impact the outcomes in this study.

Works Cited

Abraham, C. M. "A Brief Historical Perspective on Dental Implants, Their Surface Coatings and Treatments." The Open Dentistry Journal, vol. 8, no. 1, 2014, pp. 50–55, https://doi.org/10.2174/1874210601408010050.

Atagün, Ö. S., and Ü. T. Ö. Kalyoncuoğlu. "Evaluation of Elderly Patients' Knowledge and Awareness of Dental Implant Treatments Applying to Periodontology and Prosthodontics Departments." European Journal of Geriatrics and Gerontology, vol. 5, no. 2, 2023, pp. 132–138, https://doi.org/10.4274/ejgg.galenos.2023.2022-12-5.

Bors, A., et al. "From Conventional to Smart Prosthetics: Redefining Complete Denture Therapy through Technology and Regenerative Science." Medicina, vol. 61, no. 6, 2025, p. 1104, https://doi.org/10.3390/medicina61061104.

Glenza, J. "US Trials Developed Democracies in Healthcare Costs for Older Adults." The Guardian, 4 Dec. 2024, https://www.theguardian.com/us-news/2024/dec/04/americans-healthcare-cost-global-democracies.

Nasseh, K., et al. "Benefit Design and Access to Dental Care among Seniors with Medicare Advantage Dental Benefits." JAMA Health Forum, vol. 6, no. 1, 2025, https://doi.org/10.1001/jamahealthforum.2024.5123.

Neves, A., et al. "Safeguarding Our Dentistry Heritage: A Study of the History and Conservation of Nineteenth–Twentieth Century Dentures." Heritage Science, vol. 11, no. 1, 2023, https://doi.org/10.1186/s40494-023-00989-2.

Pokorny, P. H., et al. "Occlusion for Fixed Prosthodontics: A Historical Perspective of the Gnathological Influence." The Journal of Prosthetic Dentistry, vol. 99, no. 4, 2008, pp. 299–313, https://doi.org/10.1016/s0022-3913(08)60066-9.

Prakash, J., et al. "The Status, Need, and Influence of Dental Prosthetics on Oral Health-Related Quality of Life in the Geriatric Population: An Epidemiological Survey." Cureus, 2022, https://doi.org/10.7759/cureus.27637.

Samiou, A. C., et al. "Association between the Type of Dental Prostheses and the Masticatory Performance of Institutionalized Patients: A Cross-Sectional Study." Archives of Gerontology and Geriatrics Plus, vol. 1, no. 4, 2024, p. 100094, https://doi.org/10.1016/j.aggp.2024.100094.

Taue, R., et al. "Oral Function Status of Older Patients Seeking Dental Implant Treatment." International Journal of Implant Dentistry, vol. 10, no. 1, 2024, https://doi.org/10.1186/s40729-024-00571-w.

UIC College of Dentistry. "What Older Adults Need to Know about Oral Health." UI Health, 15 Mar. 2017.

Yu, S., et al. "Impact of Oral Health Literacy on Oral Health Behaviors and Outcomes among the Older Adults: A Scoping Review." BMC Geriatrics, vol. 24, no. 1, 2024, https://doi.org/10.1186/s12877-024-05469-1.

Zhao, Y., and B. Kim. "The Effects of the Expansion of Dental Care Coverage for the Elderly." Healthcare, vol. 12, no. 19, 2024, p. 1949, https://doi.org/10.3390/healthcare12191949.

Social Comparison as a Mediator Between Social Media Use and Adolescent Mental Health
By Leanne Mathews

Abstract

Social media has become a central part of adolescents' daily communication, shaping how young people connect with peers, construct identities, and evaluate themselves in relation to others. Recent survey data indicate that adolescents spend an average of approximately 4.8 hours per day on social media, highlighting the extent to which these platforms are integrated into daily life (Rothwell, 2023). Although research has consistently identified associations between social media use and negative mental health outcomes, including anxiety, depression, body dissatisfaction, and lower self-esteem, the processes underlying these relationships remain unclear. This literature review evaluates social comparison as a mediating mechanism between social media use and adolescent mental health. Drawing on social comparison theory and empirical research, this review synthesizes studies examining social comparison processes and mental health outcomes, including self-esteem, body image, depressive symptoms, and life satisfaction. Studies were selected based on their focus on adolescent and young adult populations, psychological outcomes, and relevant social media mechanisms. Evidence suggests that social media increases opportunities for upward social comparison by exposing adolescents to idealized portrayals of appearance, relationships, and success. These comparisons are consistently associated with lower self-esteem, poorer body image, reduced life satisfaction, and increased depressive symptoms. However, social comparison alone does not fully explain the relationship between social media use and adolescent mental health. Additional factors, including passive social media use, peer feedback systems, FoMO, cyberbullying, and individual vulnerabilities, also contribute to differences in how social media affects adolescent mental health. Understanding these mechanisms can help educators promote media literacy, support families in encouraging healthier digital habits, and guide mental health professionals and technology developers in creating environments that better support adolescent psychological health.

Introduction

Social media has become one of the primary environments in which adolescents interact, present themselves, and form perceptions of their own identities. Platforms such as Instagram, TikTok, Snapchat, and Facebook provide opportunities for connection and self-expression, but they also create constant exposure to the behaviors, appearances, and achievements of peers. Recent estimates suggest that teenagers spend an average of approximately 4.8 hours per day on social media, making these platforms a substantial part of adolescents' daily lives and an important context for psychological development (Rothwell, 2023).

Understanding the relationship between social media use and adolescent mental health has become increasingly important as concerns surrounding adolescent psychological well-being have grown. Twenge et al. (2018) reported that adolescents who spent more time using digital

media were more likely to report depressive symptoms and suicide-related outcomes than peers who spent less time on these activities. Specifically, adolescents who used electronic devices for three or more hours per day were 34% more likely to report at least one suicide-related outcome, and adolescents who used social media daily were 13% more likely to report high levels of depressive symptoms than those who used social media less frequently (Twenge et al., 2018). Although these findings demonstrate meaningful associations, the authors emphasized that they do not establish causation. Similarly, Odgers and Jensen (2020) concluded that although social media use is associated with poorer mental health for some adolescents, the magnitude of these relationships varies considerably across individuals, suggesting that social media alone cannot account for changes in adolescent mental health. Instead, researchers have increasingly focused on the psychological processes that shape how adolescents interpret and respond to social media content, including the ways these experiences influence self-evaluation and emotional well-being.

One of the most prominent explanations is social comparison, which may function as a mediating variable between social media use and adolescent mental health. According to social comparison theory (Festinger, 1954), individuals evaluate their abilities, characteristics, and experiences by comparing themselves with others. Social media intensifies these comparisons by creating environments where users often display carefully selected and idealized aspects of their lives. For adolescents, who are already navigating identity development and possess a heightened sensitivity to peer evaluation, frequent exposure to these comparisons may influence self-esteem, body image, depressive symptoms, and life satisfaction.

Understanding social comparison as a mediator is especially relevant because it helps explain why social media affects adolescents differently. While some individuals experience connection and social support online, others engage in frequent upward comparisons that negatively influence their self-perceptions. This review examines the extent to which social comparison explains the relationship between social media use and adolescent mental health. Specifically, it synthesizes research examining outcomes including self-esteem, body image, depressive symptoms, and life satisfaction while also considering related processes such as fear of missing out (FoMO), passive social media use, and peer feedback systems. The central question is not whether social comparison contributes to adolescent mental health outcomes, but the extent to which it explains the relationship between social media use and these outcomes. Vogel et al. (2014) found that social comparison tendencies partially explained the relationship between social media use and lower self-esteem, suggesting that comparison accounts for some, but not all, of social media's psychological effects.

Method

This paper is a narrative literature review synthesizing empirical and theoretical research examining the relationship between social media use, social comparison, and adolescent mental health. The purpose of this review is to evaluate the extent to which social comparison explains the connection between social media use and psychological outcomes.

Literature was identified through Google Scholar using combinations of search terms including “social media,” “social comparison,” “adolescent mental health,” “body image,” “self-esteem,” “depression,” “life satisfaction,” “fear of missing out (FoMO),” “passive social media use,” and “social networking.” Priority was given to peer-reviewed studies published between 2010 and 2024 to ensure that the review reflected contemporary social media platforms and adolescent digital behaviors, while foundational theoretical work (Festinger, 1954) was included because of its relevance to the conceptual framework.

Sources were selected based on their relevance to adolescent and young adult populations, examination of social comparison processes, and investigation of mental health outcomes including self-esteem, body image, depressive symptoms, and life satisfaction. Studies focusing on unrelated populations, outcomes outside of psychological health, or social media effects without consideration of relevant psychological mechanisms were excluded. Priority was given to peer-reviewed empirical studies, longitudinal research, and meta-analyses to incorporate evidence from multiple research methodologies.

The final review included twelve primary sources: one foundational theoretical paper (Festinger, 1954); four primarily correlational empirical studies (Chou & Edge, 2012; Vogel et al., 2014; Fardouly et al., 2015; Nesi & Prinstein, 2015); two longitudinal studies (Frison & Eggermont, 2016, 2017); one meta-analysis (Holland & Tiggemann, 2016); two review articles examining adolescent mental health and digital media use (Odgers & Jensen, 2020; Twenge et al., 2018); one review of social networking sites and subjective well-being (Verduyn et al., 2017); and one study examining Fear of Missing Out (Przybylski et al., 2013). Together, these sources provide theoretical, correlational, longitudinal, meta-analytic, and review-based perspectives on social comparison and adolescent mental health.

Studies were analyzed thematically and organized around four areas: (1) social comparison and body image, (2) social comparison and depressive symptoms and life satisfaction, (3) limitations of social comparison as an explanation, and (4) additional mechanisms influencing adolescent mental health, including FoMO and passive social media use. As a narrative review, no formal statistical analysis was conducted; findings were synthesized based on reported relationships, effect sizes, and researchers' conclusions across studies.

Social Comparison Theory as a Framework

Festinger’s (1954) social comparison theory proposes that individuals evaluate their abilities, characteristics, and experiences by comparing themselves to others, particularly when objective standards for self-evaluation are unavailable. While social comparison can provide motivation and information, upward comparisons, in which individuals compare themselves to people perceived as more successful or desirable, may contribute to negative self-evaluations when individuals perceive themselves as falling short of these standards (Festinger, 1954). Although Festinger developed this theory decades before the rise of social media, its principles

remain highly relevant in today's digital environment, where opportunities for comparison are more frequent, immediate, and difficult to avoid.

Social media has intensified these comparison processes by increasing adolescents' exposure to potential comparison targets. Unlike many offline interactions, social media provides continuous access to carefully curated representations of peers, influencers, and public figures. Users often present selectively positive aspects of their lives, emphasizing achievements, experiences, and social interactions while minimizing everyday challenges. In addition, recommendation algorithms and engagement-driven platform features can repeatedly promote highly visible content, increasing adolescents' exposure to idealized representations of success and social status. As a result, social media may create frequent opportunities for upward comparison in areas closely tied to identity development and self-evaluation (Vogel et al., 2014; Chou & Edge, 2012).

This process is particularly relevant during adolescence, a developmental period characterized by identity formation, heightened sensitivity to peer evaluation, and the development of self-concept. While adolescents have always compared themselves to peers in offline settings, algorithmically curated social media environments make these comparisons more persistent and personalized. Research suggests that social comparison may function as a key mechanism linking social media use to adolescent mental health outcomes by influencing how individuals evaluate themselves in relation to others (Nesi & Prinstein, 2015; Verduyn et al., 2017). For example, Vogel et al. (2014) found that social comparison tendencies significantly predicted lower self-esteem, with comparison partially explaining the relationship between social media use and self-evaluations. This pattern implies that social media influences psychological outcomes not only through exposure to content itself, but also through the cognitive processes adolescents use to interpret and evaluate that content.

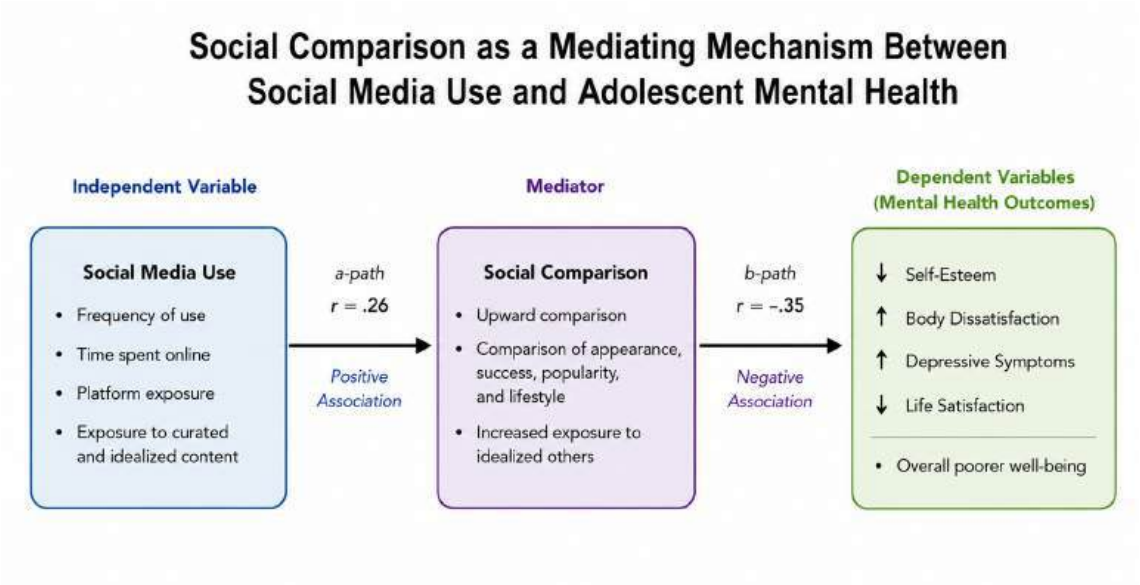


Fig. 1. Conceptual model of social comparison as a mediating mechanism between social media use and adolescent mental health outcomes.

The proposed model is supported by previous research demonstrating significant associations between social media use, social comparison processes, and adolescent mental health outcomes. The solid line representing the relationship between social media use and social comparison reflects evidence that increased exposure to social networking platforms creates more opportunities for comparison with others. For example, Vogel et al. (2014) found that individuals who used Facebook more frequently reported greater exposure to upward social comparisons, which subsequently predicted lower self-esteem. Similarly, Lee (2014) found that greater Facebook use intensity was positively associated with increased social comparison frequency, suggesting that social media environments encourage users to evaluate themselves relative to others.

The solid line connecting social comparison to mental health outcomes is supported by research showing that upward comparisons on social media are associated with negative psychological outcomes. Vogel et al. (2014) demonstrated that exposure to upward comparison information on social media was linked to reduced self-esteem and more negative self-evaluations. Additionally, Fardouly et al. (2015) and Holland and Tiggemann (2016) found that appearance-related comparisons on social networking platforms were associated with greater body dissatisfaction, supporting the relationship between comparison processes and poorer well-being. Nesi and Prinstein (2015) further showed that adolescents who engaged in more frequent social comparison on social media reported higher levels of depressive symptoms, reinforcing the connection between comparison behaviors and psychological distress.

Together, the statistical strength between these variables supports the use of solid lines rather than dotted lines because existing literature consistently identifies social comparison as a mechanism linking social media exposure with negative mental health outcomes. Although individual differences and contextual factors may influence the strength of these relationships, the current evidence supports social comparison as a significant mediating pathway between social media use and adolescent well-being.

Social Comparison and Body Image

Body image represents one of the most studied pathways through which social comparison influences adolescent mental health. Because many social media platforms prioritize visual content, adolescents are frequently exposed to idealized portrayals of appearance from peers, influencers, and public figures. These images may shape perceptions of attractiveness and social value, making appearance-based comparison particularly influential during adolescence. Platform features such as beauty filters, photo-editing tools, and curated visual content may further reinforce unrealistic appearance standards by presenting digitally altered images as typical or desirable.

Fardouly et al. (2015) found that Facebook use was associated with increased appearance comparison, which in turn predicted greater body dissatisfaction among young women. Their mediation analysis suggested that appearance comparison functioned as an explanatory pathway between Facebook use and body dissatisfaction, supporting the idea that social media affects

body image through psychological processes rather than direct exposure alone. These findings demonstrate that the effects of social media depend not only on what adolescents see online, but also on how they interpret and evaluate themselves in relation to those images.

Holland and Tiggemann's (2016) meta-analysis further supported this relationship by examining 67 studies involving more than 40,000 participants and identifying consistent associations between social networking site use and body image concerns. Their findings identified appearance comparison, internalization of appearance ideals, and body surveillance as recurring mechanisms connecting social media use to negative body image outcomes. Although effect sizes varied across studies, the consistency of findings suggests that appearance-based comparison represents an important pathway linking social media exposure with body dissatisfaction.

Together, these findings suggest that social media influences body image primarily by increasing opportunities for appearance-based comparison. Body image therefore represents one of the clearest examples of how social comparison functions as a pathway connecting online experiences with adolescent psychological outcomes.

Social Comparison and Psychological Outcomes

Although body image represents one of the most studied outcomes associated with social comparison, comparison processes also influence broader psychological outcomes, including depressive symptoms and life satisfaction. These outcomes reflect adolescents' overall evaluations of their lives, emotional well-being, and perceptions of personal success rather than appearance concerns alone. Research examining these outcomes provides further evidence that social comparison functions as a pathway through which social media use may influence adolescent mental health.

Chou and Edge (2012) found that heavier Facebook users were more likely to perceive others as happier and more successful, which was associated with lower evaluations of their own lives. Because social media often presents selective and idealized portrayals of others' experiences, adolescents may compare their own lives against unrealistic standards and conclude that they are less successful or fulfilled. These findings demonstrate that social media may influence psychological outcomes not simply through exposure to content, but through the interpretations and comparisons individuals make based on that content.

Longitudinal research provides additional evidence that social comparison may contribute to changes in psychological outcomes over time. Frison and Eggermont (2016) examined adolescents' social networking site use and life satisfaction across multiple measurement periods and found that peer comparison on social networking sites predicted declines in life satisfaction even after controlling for baseline levels of life satisfaction. Because the study assessed changes in participants' well-being rather than only differences between individuals at a single point in time, the findings provide stronger evidence that comparison processes may contribute to later psychological outcomes. However, the study does not establish

that social comparison alone causes declines in life satisfaction, as other individual and contextual factors may also influence these changes.

Research also suggests that upward comparison is associated with depressive symptoms. Frison and Eggermont (2017) found that Instagram use involving upward social comparison was associated with increased depressive symptoms among adolescents. These findings indicate that comparing oneself to individuals perceived as more successful, attractive, or socially accepted may negatively influence emotional well-being by reinforcing unfavorable self-evaluations. Together, these studies suggest that social comparison contributes to adolescent mental health outcomes by shaping how young people evaluate their own experiences, achievements, and social standing in relation to others.

However, these findings also demonstrate that social comparison explains only part of the relationship between social media use and mental health. The effects of comparison depend on factors such as the type of content viewed, the way adolescents engage with social media, and individual differences in vulnerability. Therefore, while social comparison represents an important psychological pathway connecting social media use with depressive symptoms and life satisfaction, additional mechanisms are necessary to fully understand variations in adolescent mental health outcomes.

Additional Mechanisms: FoMO, Passive Use, and Peer Dynamics

While social comparison represents an important pathway connecting social media use and adolescent mental health outcomes, it does not operate independently. Other psychological and social processes influence how adolescents engage with digital environments and how strongly social media experiences affect well-being.

Fear of Missing Out (FoMO), defined as anxiety that others are experiencing rewarding events without oneself, is closely connected to social media engagement. Przybylski et al. (2013) found that higher levels of FoMO were associated with increased social media use, greater anxiety, and lower life satisfaction. Because social media provides continuous access to peers' activities and experiences, it may intensify FoMO by increasing adolescents' awareness of events and relationships from which they feel excluded.

Usage style is another important factor influencing psychological outcomes. Verduyn et al. (2017) found that passive browsing is associated with greater envy and poorer psychological outcomes, whereas active communication tends to produce more neutral or positive effects. Passive consumption may increase exposure to idealized portrayals of others without providing meaningful interaction, creating conditions where upward comparison becomes more likely.

Social media also transforms peer evaluation by making social feedback highly visible and measurable. Nesi and Prinstein (2015) emphasize that indicators such as likes, comments, and follower counts can increase adolescents' sensitivity to peer approval and create additional sources of self-evaluation. Although these numerical measures do not accurately represent social acceptance or personal value, adolescents may interpret them as signals of popularity or status, reinforcing concerns about belonging and social standing.

Additionally, broader trends in adolescent mental health cannot be explained by social comparison alone. Twenge et al. (2018) identified increases in depressive symptoms occurring alongside rising digital media use but emphasized that multiple social and environmental factors may contribute to these patterns. These findings support the broader perspective proposed in this review: social comparison represents a key pathway linking social media use to adolescent mental health outcomes, but its effects occur alongside other psychological, behavioral, and contextual influences.

Discussion

The existing literature supports social comparison as a central mechanism connecting social media use and adolescent mental health outcomes. By increasing exposure to idealized content and comparison targets, social media creates conditions in which upward comparison becomes more frequent, influencing self-esteem, body image, life satisfaction, and depressive symptoms.

Some of the strongest evidence appears in appearance-related research, where comparison processes consistently explain part of the relationship between social media exposure and body dissatisfaction. For example, Holland and Tiggemann's (2016) meta-analysis demonstrated that appearance comparison represents a consistent pathway connecting social media use and body image concerns. Similar patterns appear in research on broader psychological outcomes, suggesting that comparison affects not only how adolescents view their appearance but also how they evaluate their overall success, relationships, and personal experiences.

However, social comparison is only one component of a larger relationship. Although social comparison demonstrates consistent associations with adolescent mental health outcomes, the magnitude of these relationships varies across studies and depends on factors such as individual vulnerability, platform use patterns, and social context. Factors such as FoMO, passive use, peer feedback systems, and individual differences indicate that adolescent mental health outcomes result from multiple interacting processes.

Therefore, social comparison should be understood as a major pathway through which social media influences adolescent mental health, rather than a single explanation for all psychological effects.

Limitations of Social Comparison

Although social comparison is a significant mechanism connecting social media use and adolescent mental health outcomes, it does not fully explain this relationship. Research demonstrates consistent associations between social comparison and psychological outcomes; however, comparison represents only one factor within a larger network of influences that shape adolescent mental health.

Methodological limitations also affect the interpretation of these findings. Many studies rely on self-reported measures of social media use and mental health outcomes, which may

introduce reporting biases or inaccuracies. Additionally, although longitudinal research provides stronger evidence than cross-sectional studies, many findings cannot establish definitive causal relationships between social media experiences, social comparison, and mental health outcomes. The reliance on correlational designs remains a major limitation because significant associations do not demonstrate whether social media use increases comparison behaviors, whether vulnerable adolescents are more likely to engage in comparison, or whether both processes occur simultaneously.

Individual differences further complicate this relationship. Adolescents with lower self-esteem, greater sensitivity to peer evaluation, or existing mental health concerns may be more likely to engage in comparison behaviors, suggesting that social comparison may function both as a contributor to and a consequence of negative psychological outcomes (Nesi & Prinstein, 2015; Odgers & Jensen, 2020). These limitations demonstrate why social comparison should be viewed as one mechanism within a broader system rather than a complete explanation.

Future research should continue examining how individual characteristics, patterns of social media engagement, and specific platform features influence the effects of social comparison. Future studies using longitudinal, experimental, and ecological momentary assessment methods may help clarify the direction and strength of these relationships by examining how comparison experiences unfold in adolescents' everyday lives. Additionally, research should explore how different social media algorithms can be designed to promote supportive interactions and reduce harmful comparison processes.

Conclusion

Research suggests that social comparison plays a significant mediating role between social media use and adolescent mental health outcomes. By increasing exposure to idealized portrayals of appearance, success, and social experiences, social media creates more opportunities for upward comparison, which is associated with lower self-esteem, poorer body image, reduced life satisfaction, and increased depressive symptoms.

However, the relationship between social media and adolescent mental health is complex. Additional factors, including FoMO, passive consumption, peer feedback systems, and individual differences, influence how adolescents interpret and respond to online environments. Social media is not inherently harmful; rather, its psychological effects depend on patterns of engagement and the meaning adolescents assign to their online experiences.

Understanding social comparison is increasingly important in today's digital society, where social media has become deeply integrated into adolescent communication, identity development, and peer relationships. These findings have practical implications beyond research. Educators can incorporate media literacy and critical thinking skills into classrooms to help students recognize unrealistic online portrayals, understand algorithmic curation, and engage with social media more thoughtfully. Families can encourage open conversations about online experiences, promote balanced technology habits, and model healthy digital behaviors. Mental health professionals can incorporate discussions of social comparison into prevention and

intervention efforts by helping adolescents identify harmful comparison patterns and develop healthier self-evaluations. Technology developers can reduce opportunities for harmful comparison by designing platforms that promote meaningful interaction over visible popularity metrics and implementing features that increase transparency, such as labeling digitally altered images or indicating when appearance-enhancing filters have been applied. Ultimately, understanding social comparison provides insight into not only why social media affects adolescent mental health, but also how evidence-based education, clinical practice, and platform design can better support adolescents in an increasingly digital world.

Works Cited

Chou, Hui-Tzu Grace, and Nicholas Edge. "'They Are Happier and Having Better Lives than I Am': The Impact of Using Facebook on Perceptions of Others' Lives." Cyberpsychology, Behavior, and Social Networking, vol. 15, no. 2, 2012, pp. 117–21, https://doi.org/10.1089/cyber.2011.0324.

Fardouly, Jasmine, et al. "Social Comparisons on Social Media: The Impact of Facebook on Young Women's Body Image Concerns and Mood." Body Image, vol. 13, no. 1, 2015, pp. 38–45, https://doi.org/https://doi.org/10.1016/j.bodyim.2014.12.002.

Festinger, Leon. "A Theory of Social Comparison Processes." Human Relations, vol. 7, no. 2, May 1954, pp. 117–40, https://doi.org/10.1177/001872675400700202.

Frison, Eline, and Steven Eggermont. "Exploring the Relationships Between Different Types of Facebook Use, Perceived Online Social Support, and Adolescents' Depressed Mood." Social Science Computer Review, vol. 34, no. 2, Jan. 2016, pp. 153–71, https://doi.org/10.1177/0894439314567449.

Frison, Eline, and Steven Eggermont. "Browsing, Posting, and Liking on Instagram: The Reciprocal Relationships Between Different Types of Instagram Use and Adolescents' Depressed Mood." Cyberpsychology, Behavior and Social Networking, vol. 20, no. 10, 2017, pp. 603–09, https://doi.org/10.1089/cyber.2017.0156.

Holland, Grace, and Marika Tiggeman. "A Systematic Review of the Impact of the Use of Social Networking Sites on Body Image and Disordered Eating Outcomes." Body Image, vol. 17, no. 1, June 2016, pp. 100–10, https://doi.org/10.1016/j.bodyim.2016.02.008.

Nesi, Jacqueline, and Mitchell J. Prinstein. "Using Social Media for Social Comparison and Feedback-Seeking: Gender and Popularity Moderate Associations with Depressive Symptoms." Journal of Abnormal Child Psychology, vol. 43, no. 8, Apr. 2015, pp. 1427–38, https://doi.org/10.1007/s10802-015-0020-0.

Odgers, Candice L., and Michaeline R. Jensen. "Annual Research Review: Adolescent Mental Health in the Digital Age: Facts, Fears, and Future Directions." Journal of Child Psychology and Psychiatry, vol. 61, no. 3, Jan. 2020, pp. 336–48, https://doi.org/10.1111/jcpp.13190.

Przybylski, Andrew K., et al. "Motivational, Emotional, and Behavioral Correlates of Fear of Missing Out." Computers in Human Behavior, vol. 29, no. 4, July 2013, pp. 1841–48, https://doi.org/https://doi.org/10.1016/j.chb.2013.02.014.

Twenge, Jean M., et al. "Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates among U.S. Adolescents after 2010 and Links to Increased New Media Screen Time." Clinical Psychological Science, vol. 6, no. 1, Nov. 2017, pp. 3–17, https://doi.org/10.1177/2167702617723376.

Verduyn, Philippe, et al. "Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review." Social Issues and Policy Review, vol. 11, no. 1, Jan. 2017, pp. 274–302, https://doi.org/https://doi.org/10.1111/sipr.12033.

Vogel, Erin A., et al. “Social Comparison, Social Media, and Self-Esteem..” Psychology of Popular Media Culture, vol. 3, no. 4, 2014, pp. 206–22, https://doi.org/https://doi.org/10.1037/ppm0000047.

Rothwell, Jonathan. “Teens Spend Average of 4.8 Hours on Social Media Per Day.” Gallup, Gallup, Inc., 13 Oct. 2023, https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx.

Jump or Drift? Comparative Empirical Analysis of Black-Scholes and Merton Jump-Diffusion Option Pricing Across Market Regimes

By Sachin Srivatsav Sathyanarayanan

Abstract

This study looks at how the Black-Scholes and Merton jump-diffusion models perform by analyzing more than 1.2 million SPY end-of-day call option quotes from January 2020 to December 2022. This time frame includes the COVID-19 market crash and the recovery that followed. Instead of matching the models directly to option prices, both are fitted to historical SPY log returns. In the Black-Scholes model, volatility is measured as the annualized standard deviation of returns. The Merton model uses four parameters, which are estimated using a method that keeps the results within reasonable limits. This is important because an unconstrained method yields an unrealistic estimate of almost seventy jumps per year, showing how hard it can be to separate volatility, jump frequency, and jump size when working with daily data. The study identifies four different market regimes around the COVID-19 Crash and checks how each model's pricing error changes overall and based on moneyness, time to maturity, and regime. The jump-diffusion model lowers the mean absolute error by about 12.5 percent compared to Black-Scholes, especially for out-of-the-money and at-the-money options, and does even better for options with longer expirations. However, when looking at the entire sample, the jump-diffusion model does worse than Black-Scholes during the Crash regime. Adjusting the models for each regime helps the jump-diffusion model do better, but both become less accurate overall. These results suggest that the volatility seen in historical data does not always match the forward-looking volatility that option prices reflect during these times.

Introduction

Determining appropriate option pricing remains a significant challenge in quantitative finance. Although the Black-Scholes and Merton models were introduced nearly fifty years ago, they continue to be the predominant method for option valuation (Black and Scholes). The model is widely favored due to its computational simplicity. Given the underlying asset price, strike price, time to expiration, risk-free interest rate, and a single volatility parameter, the model produces arbitrage-free prices efficiently. Additionally, the formula can be inverted to derive implied volatility from observed option prices. However, this simplicity relies on several strong assumptions that frequently do not align with actual market behavior. The Black-Scholes model presumes continuous price movements and constant volatility, implying log-normally distributed returns and the absence of abrupt price jumps. Empirical evidence, however, indicates that asset returns often exhibit fat tails, periods of changing volatility, and occasional large jumps that the model fails to capture (Cont). The implied volatility smile is a prominent manifestation of these limitations. When practitioners extract implied volatility from observed option prices using the Black-Scholes formula, the resulting values vary with strike price and expiration, contradicting the model's assumption of constant volatility.

Merton introduced a significant enhancement to earlier models by incorporating a jump process alongside the standard continuous price changes. This modification enables the model to capture both small, regular fluctuations and sudden, substantial jumps in asset prices. Despite this additional feature, the model retains a pricing formula that extends the Black-Scholes framework. The jump component introduces parameters such as jump frequency, average jump size, and jump size variability. These parameters allow the model to better replicate the skewness and kurtosis observed in empirical market data. Nevertheless, the practical effectiveness of this increased flexibility depends on actual market behavior rather than theoretical considerations alone.

The COVID-19 market Crash in early 2020 provides a unique opportunity to examine market responses during extreme events. Within a few weeks, the S&P 500 declined by approximately one-third, and volatility reached levels not observed since the 2008 financial crisis. Subsequently, the market began to recover and demonstrated increased stability throughout 2021 and 2022. As the market transitioned through distinct phases—calm, crisis, recovery, and a new steady state—it became possible to evaluate the performance of various financial models using empirical data rather than relying solely on simulations.

This paper addresses three primary research questions. First, it examines whether the Merton jump-diffusion model produces more accurate option prices than the Black-Scholes model when both are calibrated using historical returns rather than option prices. Second, it investigates the specific conditions—such as varying levels of moneyness, time to maturity, and market environment—under which any advantage of the Merton model emerges, intensifies, or diminishes. Third, it evaluates the impact of calibration methodology on model performance, particularly when comparing the use of a single parameter set across all data to recalibration within distinct market regimes. The principal contribution of this study is to answer these questions using a comprehensive database of 1,226,479 quoted call options, which substantially exceeds the sample sizes typically employed in option pricing research. The findings further demonstrate that the effectiveness of jump-diffusion pricing is highly contingent on moneyness, maturity, and the interaction between historical calibration and the forward-looking nature of option prices.

Theoretical Framework

The Black-Scholes model assumes that the price of the underlying asset, denoted as S, evolves according to a geometric Brownian motion process under the physical measure:

$$dS = \mu S\, dt + \sigma S\, dW$$

where μ denotes the expected return at any given time, σ represents constant volatility, and W denotes a standard Wiener process. The model is based on several key assumptions: trading occurs continuously and without friction, implying the absence of transaction costs or taxes; no dividends are paid during the option's lifetime; and the risk-free interest rate r is constant and

known. Additionally, the model assumes the absence of arbitrage opportunities. Under these conditions, and by applying the risk-neutral measure so that the drift term becomes r, a closed-form solution can be derived for the price of a European call option with strike price K and time to expiration T:

$$C = S\Phi(d_1) - Ke^{-rT}\Phi(d_2)$$
$$d_1 = [\ln(S/K) + (r + \sigma^2/2)T] / (\sigma\sqrt{T}), \quad d_2 = d_1 - \sigma\sqrt{T}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The variables are defined as follows: S represents the current price of the underlying asset, K is the strike price, T is the time remaining until expiration (measured in years), r is the annual risk-free interest rate, and σ is the annual volatility of the asset's log returns. Within this model, σ is the sole parameter that must be estimated or obtained from market data. Because the model assumes that σ remains constant across all strikes and expiration dates, it predicts a flat implied volatility surface. Consequently, using the market price of any option to solve for σ should yield the same value regardless of the chosen strike price or expiration date.

Merton extends the diffusion in Black-Scholes by adding a compound Poisson jump component:

$$dS/S = (\mu - \lambda k)\, dt + \sigma\, dW + dJ$$

where J denotes a compound Poisson process with intensity λ, representing the expected number of jumps per year, and independent, identically distributed jump sizes. Each jump multiplies the underlying price by e^Y, where Y follows a normal distribution with mean μ_J and standard deviation σ_J. The compensator $k = E[e^Y] - 1 = \exp(\mu_J + \sigma_J^2/2) - 1$ adjusts the drift so that the jump component has zero expected instantaneous effect on the price, isolating λ, μ_J, and σ_J as parameters that shape the higher moments of the return distribution rather than its mean. If the number of jumps, n, occurring before the option expires is known, the final log-price follows a normal distribution in which both volatility and drift are functions of n. Since the number of jumps within a time period T follows a Poisson distribution with mean λT, the option price when n is unknown is a Poisson-weighted sum of Black-Scholes prices evaluated at jump-adjusted parameters:

$$C_MJD = \Sigma\ (n=0 \text{ to } \infty)\ [e^{-\lambda T}(\lambda T)^n / n!] \cdot C_BS(S, K, T, r_n, \sigma_n)$$
$$\sigma_n = \sqrt{(\sigma^2 + n\sigma_J^2/T)}, \quad r_n = r - \lambda k + n(\mu_J + \sigma_J^2/2)/T$$

In this model, each term represents the option price under the assumption that exactly n jumps occur. Volatility is adjusted to reflect the additional risk introduced by these jumps, and the drift is modified to align with the risk-neutral rate after accounting for the impact of jumps.

Because Poisson probabilities become negligible when n significantly exceeds the average number of jumps, it is sufficient to include only the first fifty terms for the jump intensities considered in this context. Setting the jump intensity to zero reduces the formula to

the standard Black-Scholes price, indicating that Merton's model encompasses Black-Scholes as a special case. The Black-Scholes formula contains a single free parameter, resulting in one implied volatility value per contract when calibrated to an option's market price. If the model were fully accurate, this implied volatility would remain constant across all strike prices and maturities. In practice, however, equity index options exhibit a pronounced skew, where implied volatility increases as the strike price decreases relative to the underlying asset. This pattern indicates that the market anticipates returns with a higher probability of negative outcomes and fatter tails than those predicted by a log-normal distribution.

The Merton jump-diffusion model explains this phenomenon. Incorporating a jump component with a negative average jump size shifts the return distribution to the left. Allowing for variability in jump sizes further increases the thickness of both tails compared to the standard diffusion model. Consequently, options such as deep out-of-the-money puts and deep in-the-money calls are assigned higher probabilities and prices than those implied by the Black-Scholes model. Calculating implied volatility from Merton jump-diffusion prices across different strikes produces a curve rather than a flat line, replicating the general shape of the observed volatility smile, even when the model is calibrated solely to return data rather than directly to the volatility smile.

Data and Methodology

The SPY Daily End-of-Day Options Quotes dataset from Kaggle, published by kylegraupe under a CC BY-SA 4.0 license, was utilized for this study. This dataset contains option quotes for the SPDR S&P 500 ETF Trust (SPY) spanning January 2020 to December 2022. The original file comprises 3,589,079 observations and thirty-three fields, including bid and ask prices for both calls and puts, implied volatility, Greeks, days until expiration, and the distance between the strike and underlying prices. Due to the file size of 1.28 gigabytes, the data was loaded in increments of 500,000 rows. To ensure relevance, the dataset was filtered to retain contracts with seven to ninety days until expiration, non-zero bid and ask prices, and strike prices within thirty percent of the underlying price, corresponding to a moneyness range of approximately 0.70 to 1.30. Observations missing essential pricing information were excluded. After filtering, the final sample consisted of 1,226,479 observations, representing approximately 34.2 percent of the original data. The cleaned dataset covers January 2, 2020 to December 30, 2022, with an average moneyness (S/K) of 1.049 and a standard deviation of 0.131.

Observations were categorized into four market regimes according to their quote dates: Pre-Crash (before February 20, 2020; 41,459 observations), Crash (February 20 through April 30, 2020; 82,154 observations), Recovery (May 1 through December 31, 2020; 249,645 observations), and Post-Recovery (January 1, 2021 onward; 853,221 observations). Notably, the number of observations is not balanced across these periods. For instance, the Crash period spans approximately ten weeks, whereas the Post-Recovery period encompasses two full years. This discrepancy in time frames represents a limitation of the data and is addressed further in the discussion of results for each regime.

Rather than relying on option prices, both models were calibrated using the daily price history of SPY. This methodology emphasizes the ability of a model, grounded in the actual return process, to price options. I collected daily adjusted closing prices from yfinance for the period January 1, 2019 to December 31, 2022, resulting in 1,007 daily log returns. To enhance the stability of the volatility estimate, I also incorporated data from before 2020. Black-Scholes volatility was estimated by calculating the sample standard deviation of these daily log returns and annualizing it using the square root of 252 trading days. The full-sample volatility estimate was $\sigma = 0.2264$. Across different periods, volatility ranged from 0.1242 before the Crash to 0.6660 during the Crash.

Estimation of the Merton jump-diffusion parameters, including diffusion volatility (σ), jump intensity (λ), mean jump size (μ_J), and jump-volatility (σ_J), was conducted using maximum likelihood. This method models daily log-returns as a mixture of normal distributions, conditional on the number of jumps. Initially, an unconstrained optimization was performed using the Nelder-Mead simplex algorithm, resulting in a high jump intensity of approximately 69.8 jumps per year. Rather than indicating rare, large jumps, this finding suggests that the model captures numerous small jumps, which accounts for the excess kurtosis observed in daily returns. This result underscores a common identification issue: distinguishing the effects of diffusion volatility, jump intensity, and jump-size volatility is challenging when relying solely on the overall return distribution. To obtain economically meaningful parameter estimates, the estimation was repeated using a bounded differential evolution optimizer. Reasonable bounds were imposed on each parameter: jump intensity between 0.5 and 10 jumps per year, mean jump size between -0.15 and 0.05, jump volatility between 0.01 and 0.20, and diffusion volatility between 0.05 and 0.50. Calibration was performed on the full sample as well as within each regime to assess the sensitivity of the results. Throughout the analysis, a constant risk-free rate of $r = 0.02$ was applied.

Model accuracy was assessed by comparing model-generated prices to each contract's observed mid-price, defined as the average of the bid and ask prices. Three primary metrics were employed: mean absolute error (MAE), root mean squared error (RMSE), and bias, defined as the average signed error. RMSE assigns greater weight to large errors than MAE and is therefore more sensitive to extreme deviations. Bias indicates whether a model systematically overprices or underprices contracts. These accuracy metrics were computed for the entire dataset as well as within specific subgroups. The subgroups comprised five moneyness categories (ranging from deep out-of-the-money to deep-in-the-money, with thresholds at 0.90, 0.97, 1.03, and 1.10), three maturity ranges (short: 7-21 days, medium: 22-45 days, and long: 46-90 days), and four distinct market regimes.

Results

Table 1 presents the calibrated parameters for both models using the full sample, alongside the unconstrained Nelder-Mead jump-diffusion estimate for comparison. When the optimizer is constrained, the jump intensity decreases from 69.8 to 10 jumps per year, which is

the upper limit permitted by the bound. Concurrently, diffusion volatility increases from 0.124 to 0.163, and jump volatility nearly doubles from 0.021 to 0.041. Because the optimizer cannot utilize frequent jumps to explain the excess kurtosis, it compensates by increasing diffusion volatility and the magnitude of jumps.

Parameter	Black-Scholes	MJD (Nelder-Mead)	MJD (Bounded DE)
σ (diffusion vol.)	0.2264	0.1236	0.1627
λ (jumps/year)	—	69.82	10.00
μ_J (mean jump)	—	−0.0043	−0.0125
σ_J (jump vol.)	—	0.0214	0.0412

Table 1: Full-sample calibrated parameters

Table 2 presents the results of the bounded calibration for each regime individually. In all regimes, the jump intensity reaches the upper limit of 10 per year. This finding indicates that the identification problem is not limited to the full sample but also arises when the model is fitted to a return series using maximum likelihood estimation. The Crash regime is particularly notable: diffusion volatility reaches its upper bound of 0.50, jump volatility falls to its minimum value of 0.01, and the mean jump size is −0.098, which is nearly eight times greater in magnitude than in any other regime. These results indicate that during the Crash, the period was sufficiently volatile that most of the additional variance is attributed to the diffusion component, while the model still identifies a pronounced negative average jump consistent with the substantial losses observed during the Crash.

Regime	BS σ	MJD σ	MJD λ	MJD μ_J	MJD σ_J
Pre-Crash	0.1242	0.0985	10.00	−0.0131	0.0165
Crash	0.6660	0.5000	10.00	−0.0983	0.0100
Recovery	0.1910	0.1515	10.00	−0.0321	0.0100
Post-Recovery	0.1948	0.1761	10.00	−0.0112	0.0224

Table 2: Per-regime bounded calibration

Table 3 presents the pricing accuracy for all 1,226,479 call options using the complete set of calibrated parameters. The jump-diffusion model outperforms the Black-Scholes model in every category, exhibiting lower mean absolute error, lower root mean squared error, and reduced bias. For instance, the mean absolute error for the jump-diffusion model is $1.2251, representing a 12.5 percent reduction compared to the $1.3998 error observed with Black-Scholes. The bias is also substantially lower at $0.1706, relative to $0.5278 for Black-Scholes. The results indicate that the primary improvement arises from correcting a consistent pattern of overpricing, rather than merely reducing the dispersion of errors.

Model	MAE ($)	RMSE ($)	Bias ($)
Black-Scholes	1.3998	2.2024	0.5278
Merton Jump-Diffusion	1.2251	1.9600	0.1706

Table 3: Overall pricing accuracy (full-sample parameters)

Table 4 presents the changes in model performance based on whether the contract is out-of-the-money, at-the-money, or in-the-money. The jump-diffusion model demonstrates significantly improved mean absolute error (MAE) for out-of-the-money and at-the-money contracts, with reductions of $0.41 and $0.35, respectively. In contrast, for in-the-money contracts, the improvement decreases to $0.01, and for deep in-the-money contracts, the difference is negligible. This result is expected, as the price of a deep in-the-money call option is primarily determined by its intrinsic value, which both models estimate similarly. Conversely, out-of-the-money and at-the-money options derive their value from portions of the price distribution that are most influenced by the jump component.

Moneyness Bucket	Count	BS MAE	MJD MAE	Improvement
Deep OTM (<0.90)	106,641	0.4192	0.3008	+0.1184
OTM (0.90–0.97)	254,382	1.8027	1.3936	+0.4091
ATM (0.97–1.03)	270,345	2.6010	2.2507	+0.3503
ITM (1.03–1.10)	244,243	1.2595	1.2446	+0.0149
Deep ITM (>1.10)	350,868	0.5777	0.5801	−0.0024

Table 4: Pricing accuracy by moneyness (full-sample parameters)

Figure 1 presents results for a typical thirty-day expiry period, utilizing data from June 15, 2021 to July 16, 2021 and encompassing 81 contracts. Both models exhibit increasing errors that peak when the option is at-the-money; however, beyond this point, the jump-diffusion model maintains lower error levels compared to the Black-Scholes model. Figure 2 examines the same dataset with a focus on the implied volatility smile. In observed market data, implied volatility demonstrates a pronounced skew. The jump-diffusion model replicates a slope in the same direction, though less pronounced, whereas the Black-Scholes model remains flat due to its structural assumptions. Consequently, the jump-diffusion model approximates the general shape of the volatility smile but does not fully capture its magnitude in real market data when calibrated historically. This limitation will be addressed in subsequent sections.

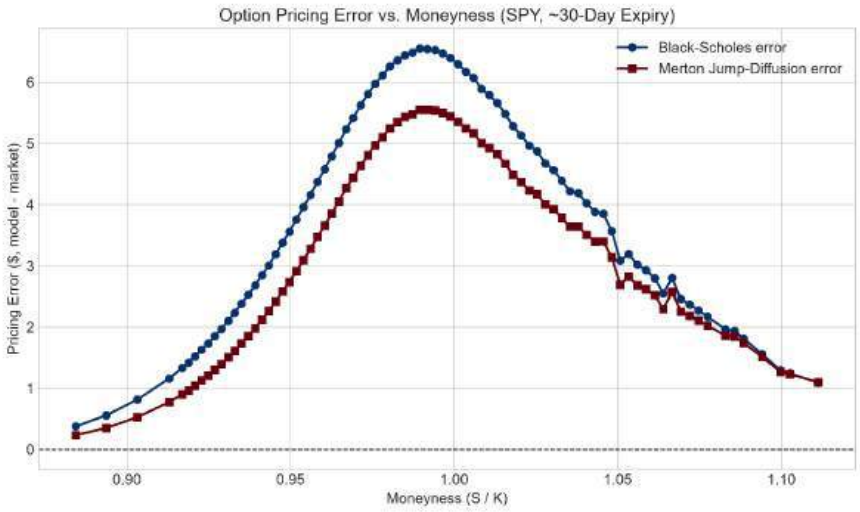


Fig 1: Option pricing error versus moneyness for SPY options with approximately thirty days until expiry (quote date 2021-06-15, expiration 2021-07-16)

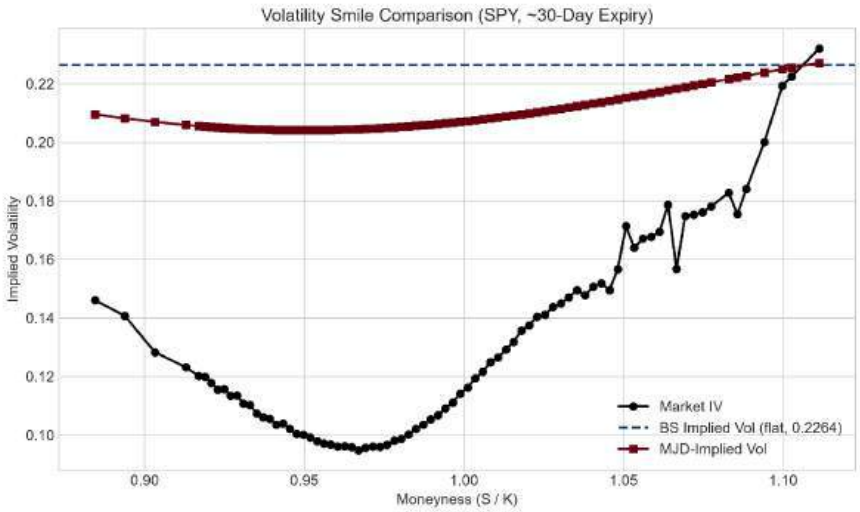


Fig 2: Implied volatility smile comparison for the same expiry slice, showing market-implied volatility, flat Black-Scholes implied volatility, and jump-diffusion implied volatility recovered by numerical inversion

Table 5 presents the variation in accuracy as a function of time remaining until expiration. The jump-diffusion model outperforms the Black-Scholes model as contract duration increases, with accuracy improvements of $0.108 for short-term contracts (7-21 days), $0.201 for medium-term contracts (22-45 days), and $0.242 for long-term contracts (46-90 days)

This trend is expected because the jump-diffusion formula incorporates the expected number of jumps, calculated as λT. Consequently, options with longer durations are more likely to experience additional jumps, resulting in price predictions that diverge further from those generated by the standard diffusion model. Figure 3 examines this trend in greater detail by

segmenting the days to expiration into twelve seven-day intervals, demonstrating that the improvement continues to increase as the time to expiration extends.

DTE Bucket	Count	BS MAE	MJD MAE	Improvement
Short (7–21 days)	454,998	0.9263	0.8184	+0.1079
Medium (22–45 days)	524,489	1.5362	1.3353	+0.2009
Long (46–90 days)	246,992	1.9821	1.7404	+0.2417

Table 5: Pricing accuracy by time to expiration

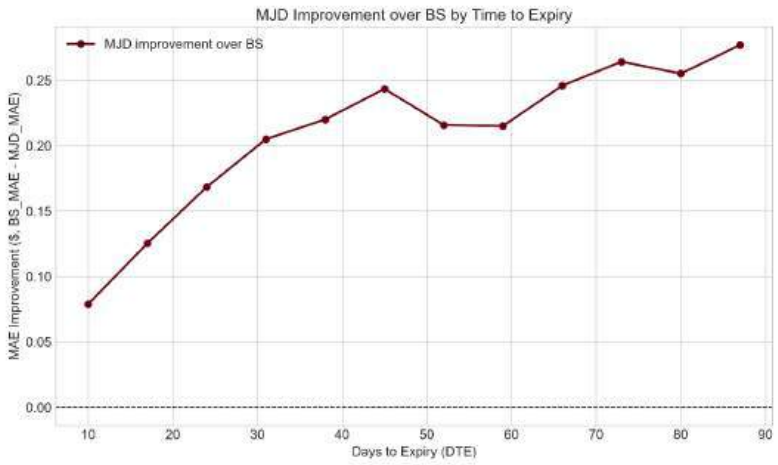


Fig 3: Jump-diffusion MAE improvement over Black-Scholes as a function of days to expiration, binned into seven-day windows

Table 6 shows the mean absolute error (MAE) for each regime using both calibration methods. When using parameters from the full sample, the jump-diffusion model does better than Black-Scholes in the Pre-Crash, Recovery, and Post-Recovery periods. However, during the Crash itself, Black-Scholes has a lower MAE ($2.77) compared to jump-diffusion ($2.91). If the models are recalibrated separately for each regime, this pattern flips in the Crash period: now, jump-diffusion performs better with an MAE of $4.13, while Black-Scholes rises to $6.23. Still, this regime–specific calibration leads to a bigger problem. Both models become much less accurate during the Crash, with their MAEs increasing substantially. For example, Black-Scholes MAE jumps from $2.77 to $6.23, and jump-diffusion goes from $2.91 to $4.13. The reason for this is that the Black-Scholes volatility used for the Crash regime, which is 0.666 and is based on actual returns during the Crash, is much higher than what option prices suggested at the time. Using the volatility, which matches history but not what traders expected, causes the model to overprice every option in that period. Figure 4 compares the two calibration methods, and Figure 5 breaks down the results by regime and moneyness using the full-sample parameters. In the Crash regime, the model errors are negative for every moneyness group.

Regime	BS MAE (full)	BS MAE (adaptive)	MJD MAE (full)	MJD MAE (adaptive)
Pre-Crash	1.7909	0.4840	1.4517	0.4524
Crash	2.7666	6.2301	2.9060	4.1296
Recovery	0.8839	0.8489	0.7809	0.9368
Post-Recovery	1.4001	1.1200	1.1822	1.0954

Table 6: Regime-level MAE, full-sample vs. regime-adaptive calibration

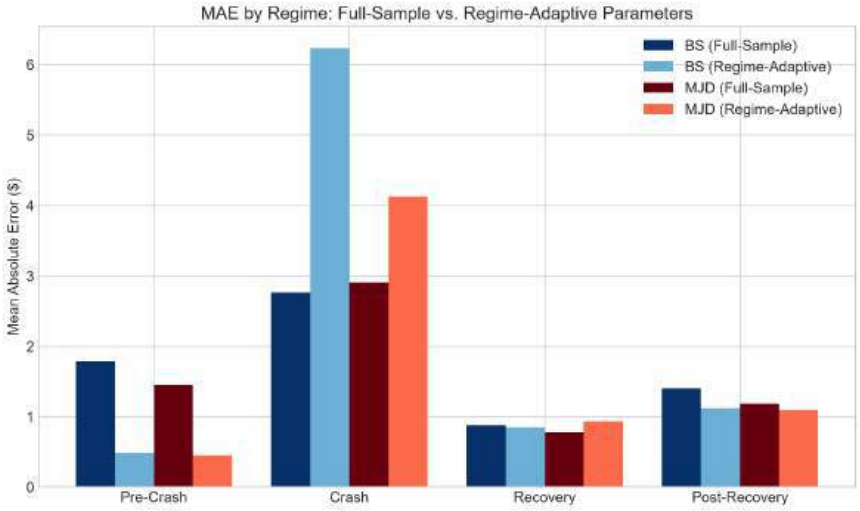


Fig 4: Mean absolute error by regime under full-sample versus regime-adaptive calibration, for both Black-Scholes and jump-diffusion models

Fig 5: Jump-diffusion MAE improvement over Black-Scholes across the regime-by-moneyness cross-tabulation (full-sample parameters). Green indicates jump-diffusion wins; red indicates Black-Scholes wins.

Discussion

The primary finding is that jump-diffusion pricing outperforms the Black-Scholes model, particularly for contracts that are out-of-the-money or at-the-money. This result is consistent with previous studies highlighting the limitations of constant volatility assumptions and the benefits of more sophisticated models. For instance, Bakshi, Cao, and Chen demonstrated that incorporating both stochastic volatility and jumps into the Black-Scholes framework reduces pricing and hedging errors, although these errors are not eliminated. The observed patterns, such as greater improvements for contracts far from the money and for those with longer maturity, further support the conclusion that no single model is optimal for all option types.

The calibration results reveal a significant challenge in applying the jump-diffusion model in practice. Parameter estimation from daily returns using maximum likelihood makes it difficult to distinguish the effects of diffusion volatility, jump intensity, and jump-size volatility. This issue arises because the overall return distribution primarily constrains the combination of these parameters rather than each parameter individually. For instance, an estimate of seventy jumps per year does not necessarily indicate that SPY experiences jumps with such frequency. Rather, this outcome demonstrates that multiple combinations of λ, σ_J, and σ can explain the observed data with similar effectiveness. The consistent tendency of the optimizer to set λ at its maximum permissible value, rather than at an intermediate level, further illustrates this identifiability problem.

The regime-adaptive results highlight a broader issue beyond parameter identification. Calibrating models to historical, physical-measure data differs fundamentally from aligning with the risk-neutral, forward-looking volatility embedded in option prices. Option prices reflect market expectation for future volatility, as well as a risk premium that increases during periods of market stress. Consequently, implied volatility during a crisis is typically lower than the realized volatility observed afterward. Implied volatility is determined before the outcome, whereas realized volatility is measured retrospectively. Calibrating the Black-Scholes model's volatility to the realized 66.6 percent annualized volatility from the Crash period results in overpricing all options within that window. The jump-diffusion model also overprices options, but to a lesser extent, as its jump component can absorb some of the additional variance. This distinction helps clarify the primary puzzle at the regime level. Regime-adaptive calibration correctly ranks the two models because the jump-diffusion model provides an alternative mechanism for accommodating the elevated volatility assigned to both models.

This concept is also relevant to the volatility smile. The jump component of the Merton Jump-Diffusion (MJD) model adjusts implied volatility appropriately across various strike prices, consistent with the theoretical expectations outlined in the Theoretical Framework section. However, when the jump distribution is estimated using historical returns, it fails to replicate the magnitude of the market's risk-neutral skew, as these distributions are derived for different purposes. For practitioners, historically calibrated jump-diffusion pricing proves most effective for out-of-the-money and at-the-money options with extended times to expiration, where the model demonstrates greater accuracy and reliability. Conversely, for deep

in-the-money options or during periods of elevated volatility, historical model parameters should not substitute for market-derived information.

Conclusion

This study examines the pricing performance of the Black-Scholes model and the Merton jump-diffusion model for SPY call options, utilizing data from over 1.2 million contracts during the COVID-19 Crash and the subsequent period. Both models are calibrated exclusively with historical returns of the underlying asset. Three principal findings emerge. First, the jump-diffusion model reduces the average pricing error by approximately 12.5 percent overall. This improvement is most pronounced for out-of-the-money and at-the-money options, and it becomes more significant as time to expiration increases. For deep in-the-money options, both models yield identical results, as pricing is primarily determined by intrinsic value. Second, calibration of the jump-diffusion model via maximum likelihood estimation exhibits instability. Without constraints, the optimizer selects an unrealistically high jump intensity, and even with imposed limits, it consistently reaches the upper bound. This suggests that distinguishing between standard volatility and jumps is challenging when relying solely on aggregate returns. Third, adapting the models to different market regimes enhances the jump-diffusion model's performance during the Crash, but also reduces the overall accuracy of both models in that period. This underscores a discrepancy between historical volatility and the risk-neutral volatility reflected in options prices.

Several important limitations should be considered when interpreting these conclusions. First, both models were calibrated exclusively using historical returns rather than actual option prices, so the results primarily illustrate the behavior of the jump-diffusion structure under this calibration approach. Additionally, the analysis focused solely on SPY and employed a European-style formula to price call options, even though most equity options in the market are American-style and allow for early exercise, which is not captured by the formulas used here. The risk-free rate was held constant at two percent throughout the three-year period, although actual short-term rates fluctuated during this time. Future research could enhance this work by calibrating the models directly to option prices to assess the extent to which discrepancies arise from differences between realized and implied volatility. Further comparisons with stochastic volatility models, such as the Heston model, or with models that combine stochastic volatility and jumps, could help determine whether the observed weaknesses are primarily attributed to the constant-volatility assumption. Finally, employing a combination of Black-Scholes and jump-diffusion prices that varies according to moneyness and maturity may leverage the performance differences identified in this study.

Works Cited

Bakshi, Gurdip, Charles Cao, and Zhiwu Chen. "Empirical Performance of Alternative Option Pricing Models." Journal of Finance 52, no. 5 (1997): 2003-2049.

Black, Fischer, and Myron Scholes. "The Pricing of Options and Corporate Liabilities." Journal of Political Economy 81, no. 3 (1973): 637-654.

Cont, Rama. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues." Quantitative Finance 1, no. 2 (2001): 223-236.

kylegraupe. "SPY Daily EOD Options Quotes 2020–2022." Kaggle. Accessed 2025. https://www.kaggle.com/datasets/kylegraupe/spy-daily-eod-options-quotes-2020-2022.

Merton, Robert C. "Option Pricing When Underlying Stock Returns Are Discontinuous." Journal of Financial Economics 3, no. 1-2 (1976): 125-144.

Pool-Based Active Learning for Efficient AI-Assisted Lightweight CubeSat Structural Optimization By Akhil Tirumalaseti

While machine learning surrogate models offer a pathway to real-time structural optimization of spacecraft hardware, they are severely limited by data scarcity in resource-constrained development environments; implementing a pool-based active learning framework with Gaussian Process Regression bridges this gap by intelligently minimizing the number of high-fidelity finite element simulations required to identify optimal, minimum-mass bracket designs within aerospace safety constraints. The finite element method (FEM) is time consuming and computationally expensive (Hoffer et al.). Every additional geometric variable a designer wants to explore multiplies the number of simulations needed to characterize the design space, and for small teams without access to institutional compute clusters, that multiplication quickly becomes the limiting factor in how thoroughly a component can actually be optimized before a mission deadline.

CubeSats are a demanding medium to work with, since they must be both small and light. The CubeSat Design Specification sets a maximum mass of 1.33 kilograms for a 1U CubeSat ("CubeSat Design Specification"). Size-wise, they are similarly constrained, since CubeSats typically range from 1U to 12U, with each U being a $10 \times 10 \times 10$ cm³ cube, roughly the size of a Rubik's cube. However, they still need to contain a wide variety of complicated instruments, from navigation systems to power systems, and because of these constraints, every part on a CubeSat needs to be thoroughly optimized. Typically, this is done with finite element analysis (FEA), but FEA is very computationally expensive and takes a lot of time, and a single research team rarely has the luxury of running it hundreds of thousands of times over the course of a single design cycle.

A different approach is to use AI trained on simulation data to predict physical outcomes, instead of actually simulating every set of parameters. Shao et al. demonstrate exactly this, using fully connected and long-short-term-memory neural networks as FEA surrogates for stress prediction in impact-loaded plates, achieving over 92% agreement with full FEA results. In doing so, simulations that used to take months are now finished in hours, and simulations that took hours are finished in seconds, greatly reducing the cost of optimization and testing for every component on a CubeSat. Traditional prototyping methods require the construction of multiple physical prototypes, leading to material waste and increased costs (Mecanica Solutions), which is especially crucial in the aerospace research environment where CubeSats are typically utilized, given the tight budgets and expensive materials involved, where a single wasted print or machining run can consume a meaningful fraction of a semester's hardware budget. By instead training models on simulation data, far more combinations of parameters can be tested without the computationally taxing process of opening and closing a simulation environment for every single run, and without requiring the researcher to babysit each individual solve. This is the underlying premise this paper investigates directly: not simply whether a surrogate model can be trained on CubeSat structural data, which prior work has already shown is possible, but whether

an active learning framework can make that training meaningfully cheaper than passive sampling for a bracket-scale component under a realistic student research budget.

Research Question

How effectively can a pool-based active learning framework utilizing Gaussian Process Regression minimize the high-fidelity finite element simulation budget required to optimize a parametric star tracker mounting bracket for a cislunar CubeSat under launch load constraints?

Literature Review

Mission Context: Cislunar CubeSats and Star Trackers

CubeSat constellations have emerged as a cost-effective platform for persistent optical navigation and space domain awareness , missions that fundamentally depend on precision attitude determination hardware operating under simultaneous mass minimization and structural survival constraints. CubeSats such as the Jervis Autonomous Module, made by Rhea Space Activity, use star trackers for navigation based on the positions of stars, asteroids, and other cislunar spacecraft. This reduces the cost of navigation relative to radio-ranging methods that depend on ground infrastructure such as NASA's Deep Space Network. Because of this, star trackers are an essential tool, especially for those working in academic environments, given the cost savings they provide relative to ground-based radio tracking, which requires scheduled contact windows with a limited number of large dish antennas and becomes a real bottleneck as constellation sizes grow. Rocket Lab has developed a constellation class star tracker, the ST-16HV, based on its heritage ST-16RT2 design and explicitly engineered for mass manufacturing to meet the short lead time needs of commercial and government satellite constellation programs (Rocket Lab), underscoring that star tracker hardware itself is no longer the primary bottleneck for constellation scale CubeSat missions; the supporting structural hardware, including the mounting bracket studied here, increasingly is. Accordingly, great care has to be taken to ensure that these star trackers are as optimized as possible. To be useful, however, star trackers must be mounted with a high level of accuracy, and recalibration becomes necessary after a sensor experiences vibration or structural deformation , exactly the kind of disturbance a mounting bracket is designed to resist during launch (Ma et al.). Therefore, having a star tracker mount that is both as light as possible and strong enough to withstand the forces and acceleration during the mission is essential to the mission. Because of this, finding optimal dimensions for the bracket is very important, and in turn, active learning frameworks can be very useful in cutting down the cost of finding them. This was precisely what I did in this paper: CADing a star tracker mount, and then using FEA to train a pool-based active learning model. The specific hardware target chosen for this study, detailed fully in the Methodology section, reflects a real commercial sensor footprint rather than an arbitrary or simplified test geometry, which keeps the resulting design space grounded in an actual mission relevant constraint set.

The Finite Element Analysis Bottleneck

While FEA has become the standard method for testing CubeSat structural components, it runs into serious problems the moment you try to use it to explore a large design space rather than just validate a single design. For a straightforward check , does this specific bracket survive a 15g axial load , FEA works well. It is accurate, well-documented, and tools like ANSYS have made it accessible even to student researchers working on tight budgets (Barsoum et al.). The problem is that a single FEA run only tells you about that one specific configuration of dimensions. The moment you start asking a more practically useful question , what is the best combination of wall thickness, riser height, gusset geometry, and lightening hole size for minimizing mass while keeping the margin of safety above zero , you need to run a separate simulation for every combination you want to test. For a bracket with eight coupled geometric parameters, even a modest sampling of five discrete values per parameter would require 5^8, or roughly 390,000 individual solves. At just two minutes per solve on a standard laptop, that works out to over 500 days of continuous computation. This is not a theoretical concern; it is a wall that immediately eliminates exhaustive simulation-based search as a realistic design strategy for any small research team, and it is the single most important practical justification for everything that follows in this paper.

What makes this bottleneck particularly relevant here is that the existing CubeSat FEA literature demonstrates exactly the pattern that leads to it. Published structural studies, including Sravanthi et al. and Barsoum et al., produce high-quality, reliable results by fixing most geometric variables and evaluating only a small set of pre-determined configurations. Sravanthi et al. confirm that Al 6061-T6 is the optimal material for internal CubeSat components on the basis of its strength-to-weight ratio and machinability when compared against Al 6063-T832 and Al 7075-T6, directly justifying the material choice used in this study. Barsoum et al. validate the structural safety workflow used here, explicitly documenting the margin of safety formula $MoS = \sigma y / (\sigma max \times SF) - 1 \geq 0$ under design loads up to 51g with a safety factor of SF = 1.5. The shared limitation is not a flaw in the methodology; it is a constraint imposed by the computational reality they all operate under: to keep the simulation count manageable, the design space has to be kept narrow. Shao et al. make this relationship between design space size and simulation cost explicit , as additional geometric parameters are varied simultaneously rather than one at a time, the number of required FEA evaluations grows combinatorially rather than linearly, so what feels like a modest increase in design freedom translates into an exponential increase in required compute. This is the connection that ties the prior literature together: the same FEA methods that Sravanthi et al. and Barsoum et al. use successfully for fixed-geometry validation become impractical the moment they are applied to the kind of multi-dimensional parametric search that actually produces optimized, minimum-mass designs. The gap is in the scalability of FEA as a search strategy. Closing that gap requires a fundamentally different approach: rather than simulating every candidate design directly, train a model on a small number of simulations that can then predict performance across thousands of unseen configurations instantly. That is where machine learning surrogates come in.

Machine Learning Surrogates for Aerospace Components

The computational bottleneck described above is not a new problem, and the aerospace engineering community has been working on solutions to it for a while. The core idea behind most of these solutions is the same: instead of running a slow physics simulation every time you want to evaluate a new design, you train a machine learning model on a smaller set of simulation results, and then use that model to predict how new, untested designs will perform in a fraction of the time. These models are called surrogates, and the research on them in aerospace structural optimization has advanced significantly in recent years. Ünlüsoy et al. provide one of the clearest examples of this approach at scale, developing a deep learning surrogate for a commercial aircraft wing structure in ANSYS and benchmarking it against several model families including Gaussian Process Regression (GPR), artificial neural networks, Random Forest (RF), and support vector regression. Their findings establish GPR and RF as the two most reliable baseline families for structural surrogate comparison, because GPR and RF consistently produced the best accuracy-to-complexity trade-off across their test cases.

What is important to understand about Ünlüsoy et al., though, is the scale they were working at, and why that scale matters directly to the design decisions made later in this study's Methodology section. A commercial aircraft wing is a massive structure with enormous geometric complexity, and their training dataset, roughly 340 high-fidelity samples, reflected that: far more simulation data than any small research team working on a CubeSat component could ever realistically generate, since generating even a fraction of that many high-fidelity structural solutions on a student ANSYS license, bound by the 32,000-node ceiling described later in this paper's Methodology section, would take months rather than days. This matters because surrogate model performance is not just about which algorithm you pick; it is also deeply tied to how much training data you have. A landing gear bracket study published the same year gets much closer to the scale relevant to this paper, systematically comparing five surrogate families, Kriging, Random Forest, XGBoost, Support Vector Regression, and a Multi-Layer Perceptron, against real parametric FEA data for an aircraft landing gear bracket across training sets ranging from 500 to 5,000 samples ("Comparative Analysis"). Its central finding is that prediction accuracy improves consistently for every one of the five model families as dataset size increases, which is a significant result for this paper : it is direct, bracket-scale evidence that surrogate accuracy is dataset-size-dependent rather than a fixed property of the algorithm, which is exactly the mechanism this paper's own Results and Discussion section later points to when explaining why GPR underperformed Random Forest at a 50-point training budget. The connection between the Ünlüsoy et al. and landing gear studies is that they together establish a spectrum: both work with substantially larger training sets, at least 340 and at minimum 500 samples respectively, than the 50 points available in this study, which is a different data regime than the one this study operates under. What neither of them addresses is what happens when the training data is severely limited, which is the actual constraint faced in this study.

Shields et al. conducted a systematic head-to-head benchmark of GPR variants against Random Forest across several experimental domains and found that anisotropic GPR models perform comparably to RF when the acquisition strategy is held constant. This is important for model selection because it means there is no clear winner between the two families by default; the right choice depends on the structure of the specific design space, on how many training points are available, and on whether the underlying input dimensions are correlated with one another in ways that might favor one family's inductive biases over the other's. Taken together, these sources build a clear case for the model-selection approach used here, benchmarking GPR, Random Forest, and a linear baseline together rather than committing to one, but they also expose the gap this study is designed to fill. None of them evaluated surrogate accuracy as a function of dataset size in any systematic way, and none of them worked with CubeSat-scale internal hardware. The question of how many FEA simulations are actually needed to build a useful surrogate for a spacecraft mounting bracket, and whether active learning can cut that number down further, remains unanswered in the literature.

Active Learning and Uncertainty Sampling Mechanics

The surrogate models described in the previous section are all trained passively , you generate a set of simulation results, feed them into the model, and whatever you happen to have sampled is what the model learns from. This works reasonably well if your simulation budget is large, because random or uniform sampling across a big dataset tends to cover the design space adequately. But when your budget is tight, as it is when each data point requires a computationally expensive FEA solve, passive sampling becomes a real liability. You end up spending simulation runs on design configurations the model already understands well, while leaving other regions of the space severely underrepresented. Active learning is the solution to this. Instead of passively accepting whatever training data you happen to generate, an active learning framework lets the model itself decide which design it most needs to see next, directing simulation resources toward the regions where uncertainty is highest and the information gain will be greatest.

The mechanism that makes this possible with GPR comes from Rasmussen and Williams, whose foundational work establishes that a Gaussian Process does not just output a single prediction for any given input; it models the entire design space as a probability distribution. For every point in that space, GPR returns both a mean prediction and a predictive standard deviation, $\sigma(x^*)$. Points close to existing training data have low σ, and points far from anything the model has seen before have high σ. This uncertainty estimate is what active learning uses as its acquisition function; at each iteration, the model simply queries whichever candidate has the highest σ value, reveals its true FEA result, adds it to the training set, and retrains the model. Over successive iterations the surrogate's predictive standard deviation across the design space should shrink fastest in exactly the regions the model previously understood least, which is the theoretical basis for expecting active learning to need fewer total simulations than random sampling to reach a comparable level of accuracy.

Giovanis et al. demonstrate this empirically for high-dimensional structural problems, showing that adaptive learning strategies that balance exploration of high-uncertainty regions against exploitation of near-optimal areas consistently outperform passive random sampling for surrogate construction. The direct quantitative evidence for how large this advantage can be comes from a study on decomposition-based optimization, which found that active-learning-guided initialization achieves a significant reduction in solution time compared to standard workflows, and that active learning specifically reduces the amount of data required to reach a given level of surrogate accuracy relative to passive selection ("Taking the Human Out"). That paper does not report a single fixed percentage that would generalize to a different problem domain, but the qualitative claim, that active querying should need less data than passive random sampling to reach comparable accuracy, is exactly the claim this study's 40-query, 40-holdout benchmarking design was built to test directly rather than assume. That study's problem domain was power-systems optimization rather than structural FEA , but the comparison it draws, active querying versus random sampling using the same surrogate model, is directly analogous to the comparison this paper performs, since both compare the same underlying surrogate architecture under two different data-acquisition policies rather than comparing two entirely different modeling approaches against each other. More recent work confirms this acquisition strategy is still the field's active area of development rather than a settled question: Sauer, Gramacy, and Higdon extend active learning to deep Gaussian process surrogates to handle the non-stationary response surfaces that a single-layer GPR, the kind used in this study, can struggle to capture, while a 2024 AIAA SciTech study applies active-learning-driven GP surrogates to failure-contour estimation for aerospace flight certification problems, using the same maximum-uncertainty query logic this paper implements ("Actively Learning Deep Gaussian Process Models"). Unlike this study, neither addresses the specific case of a severely data-constrained, single-fidelity GPR operating on a correlated, sparsely-sampled design space, which is precisely the regime this paper's results speak to.

The general framework this study relies on, letting a model choose its own next training example from a pool of unlabeled candidates rather than labeling them all upfront, is one of the oldest and most heavily cited ideas in machine learning; Settles' foundational survey names this specific setup pool-based active learning and identifies it as the most common scenario in practical applications, since real-world problems more often start with a large reservoir of unlabeled data than with the ability to synthesize new queries on demand. The pool-based architecture used in this study, in which a fixed pool of candidates is pre-computed rather than generated on the fly, follows that same established scenario and is further formalized for engineering-design contexts in "Data Efficient Surrogate Modeling for Engineering Design." Rather than generating new candidate designs as the loop runs, this study pre-computes a fixed pool of candidates, a 200-point Latin Hypercube sample, and the active learning loop queries from within that pool sequentially. This is an important design decision because it removes the need for live, real-time integration between the machine learning code and the FEA software. The true simulation results for all 200 candidates already exist in a spreadsheet before the loop

begins; the model simply selects which row to reveal next. This decoupling is what makes the workflow accessible without complex API coding, which is a nontrivial practical advantage for independent researchers. What none of these studies do, however, is apply this framework to a CubeSat internal component. The methods are proven; the application context is new.

The Core Research Gap

FEA is validated and reliable for CubeSat structural hardware, but it cannot scale to multi-parameter design search (Sravanthi et al.; Barsoum et al.). Surrogate models solve the scaling problem, and GPR-family methods in particular have been shown to be viable on bracket-type and wing-scale aerospace hardware alike, provided a sufficiently large training set is available ("Comparative Analysis"; Ünlüsoy et al.). Active learning takes this further by reducing the simulation budget needed to train a good surrogate relative to passive sampling, using GPR uncertainty as the query signal (Rasmussen and Williams; Giovanis et al.; "Taking the Human Out"). Each of these three capabilities exists in the literature. What does not exist is a study that combines all three and applies the result to the actual hardware context where the need is most acute: an internal CubeSat mounting bracket, designed under a strict simulation budget, with the goal of identifying minimum-mass configurations that survive launch loads.

The gap has three specific dimensions. First, the surrogate modeling work that comes closest to this application, the landing gear bracket study, compares surrogate families passively, without any active learning loop and without limiting the simulation budget, using training sets an order of magnitude larger than what this study had available; it assumes access to enough compute to generate a large, well-distributed training dataset without worrying about how many FEA runs that costs. Second, the active learning studies that provide the methodological backbone for this paper apply their frameworks to power-systems optimization and materials-adjacent problems, not spacecraft structural components, and do not evaluate performance in the kind of severely data-constrained regime that a student research team operating under an ANSYS student license actually faces. Third, existing CubeSat structural FEA studies validate specific fixed designs or compare a small number of material choices, but none introduce any machine learning acceleration at all (Sravanthi et al.; Barsoum et al.). The intersection of CubeSat-scale bracket hardware, GPR surrogate modeling, and pool-based active learning under a constrained simulation budget is entirely unoccupied in the published literature. Closing even one of these three gaps would be a meaningful contribution; closing all three at once, and doing so with tools and licenses that are realistically available to an undergraduate research team, is what distinguishes this study from the surrounding literature.

This study is positioned directly at that intersection. Using a 200-point Latin Hypercube dataset generated in ANSYS Mechanical 2026 R1 for a parametric Al 6061-T6 L-bracket star tracker mount, the study applies a pool-based active learning loop with GPR uncertainty acquisition and benchmarks it against passive random sampling and two additional surrogate model families across the same pool. The central question, whether uncertainty-driven query selection converges to accurate stress and mass predictions faster than random selection, and by

how much, is testable, answerable, and directly relevant to anyone trying to do structural optimization of spacecraft hardware without access to a supercomputer. To the author's knowledge, no prior study has done this for a CubeSat internal mounting component, making the present work among the first demonstrations of active-learning-driven FEA surrogate optimization at this hardware scale and data regime.

Unlike prior surrogate-modeling studies that examine spacecraft structural components broadly or that apply active learning to aerodynamic and materials problems outside spaceflight hardware, this work develops and benchmarks a pool-based, GPR-driven active learning framework specifically for a CubeSat internal mounting bracket under a strict, student-scale simulation budget, and reports honestly on where that framework did and did not outperform passive random sampling. That combination, applied to this hardware class and this data regime, is the paper's central contribution.

Methodology

This section details the complete four-phase research workflow utilized in this study: parametric computer-aided design (CAD) geometry definition, finite element analysis dataset generation via Latin Hypercube Sampling (LHS), pool-based active learning with Gaussian Process Regression (GPR), and surrogate model validation and comparative benchmarking against traditional random sampling methodologies.

Hardware Geometry and Constraints

The component under evaluation is a parametric Aluminum 6061-T6 mounting bracket explicitly designed for the AAC Hyperion ST200 sensor footprint. The hardware envelope constraints are defined as 29 × 29 × 38.1 mm with a baseline sensor mass of 42 g. The internal gusset angle is fixed at 45° to ensure an equal distribution of shear forces across the primary structural planes. This footprint was chosen because it reflects a realistic, commercially available sensor rather than a hypothetical one, which means the resulting optimized bracket geometry could, in principle, be manufactured and flight-qualified without further redesign of the sensor interface.

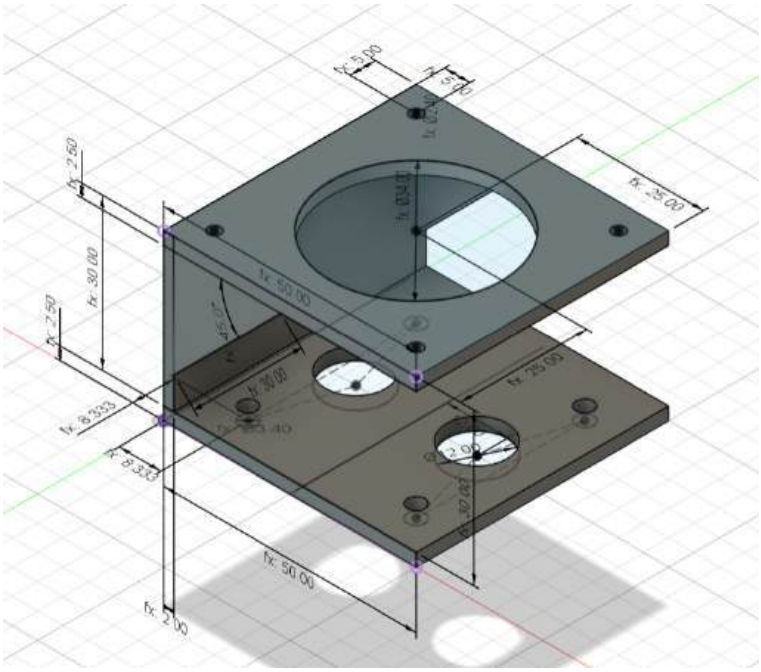


Fig. 1. parametric L-bracket CAD model with dimension callouts

Material Properties

To ensure high fidelity during the finite element analysis phases, standard material constants for Aluminum 6061-T6 were enforced. These constants include a Young's modulus (E) of 68.9 GPa, a Poisson's ratio (ν) of 0.33, a density (ρ) of 2,700 kg/m³, and a yield strength (σy) of 276 MPa. Al 6061-T6 was selected over other common aerospace aluminum alloys for its combination of machinability and strength-to-weight ratio, a choice justified in the literature review above and consistent with prior CubeSat structural studies.

Structural Load Profiles and Safety Metrics

The mounting bracket is subjected to a combined launch loading vector representing severe launch environments, yielding a total magnitude of approximately 20g. This load vector is derived from a combination of ±15g axial acceleration acting on the vertical normal axis (Z) and ±10g lateral acceleration acting simultaneously on the secondary axes (X and Y). Structural integrity and design pass/fail status are evaluated using a strict safety factor (SF) of 1.5. The structural health is quantified via the margin of safety (MoS) formula:

$$MoS = \sigma y / (\sigma max \times 1.5) - 1 \geq 0$$

where σmax represents the maximum von Mises stress extracted directly from the FEA solutions. A design point is considered viable if and only if the MoS is greater than or equal to zero. This threshold was applied uniformly across every candidate in the 200-point pool, so that the same pass/fail criterion used to evaluate the final optimized design was also used to label every point the active learning loop could potentially query. Framing viability as a single binary threshold, rather than a continuous score, also keeps the downstream interpretation of the

optimizer's output straightforward: a design either clears the bar or it does not, and the margin above that bar is a separate, secondary consideration.

CAD Parameters and Simulation Environment

Eight distinct geometric dimensions are designated as active input parameters for the optimization pool. These dimensions scale within specific programmatic bounds, while bolt hole positions scale dynamically using expressions dependent on the base width (W_base). The structural and geometric setup is defined by the following parametric values: base width (W_base, 44–60 mm), riser height (H_riser, 20–45 mm), wall thickness (t_wall, 1.5–3.5 mm), dual gusset rib thicknesses (t_rib and t_ribR, 2.0–4.0 mm), base plate thickness (t_base, 2.0–5.0 mm), top plate thickness (t_top, 1.5–3.0 mm), and lightening hole radius (r_hole, 4.0–6.0 mm). The internal gusset angle remains fixed at 45° to equally distribute shear forces, which reduces the active design space from nine to eight variables without sacrificing structural generality. Holding the gusset angle fixed also keeps the CAD model numerically stable across the full range of the other eight parameters, which matters because the Latin Hypercube sample needs every generated geometry to regenerate successfully in ANSYS without manual intervention.

Finite Element Mesh and Data Pool Generation

The parametric geometry was directly linked to ANSYS Mechanical 2026 R1 Student Edition via the Workbench Parameter Set GUI interface. To remain compliant with the student-tier licensing constraint of 32,000 nodes per solve, a uniform 1.2 mm tetrahedral element size was specified. This configuration automatically extracts mass (M) and peak von Mises stress (σmax) vectors directly to a master evaluation spreadsheet. A comprehensive 200-point database was constructed using Latin Hypercube Sampling (LHS) across the continuous eight-dimensional operational bounds to serve as the initial candidate pool. Latin Hypercube Sampling was chosen over a simple uniform grid because it guarantees that every one of the eight input dimensions is evenly represented across its full range in the resulting 200-point set, which is important for giving the surrogate model a fair, unbiased view of the design space before any active querying begins.

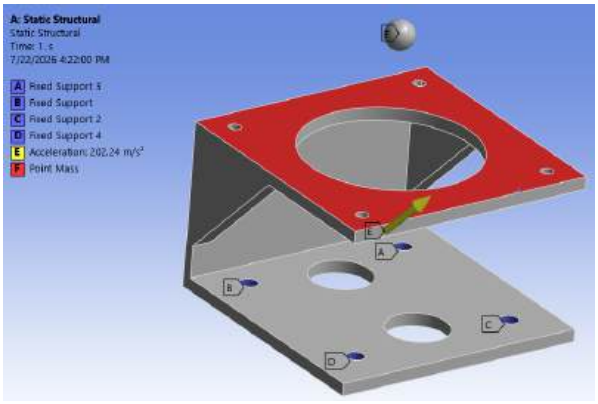


Fig. 2. ANSYS Static Structural setup , boundary conditions, fixed supports, and point mass load

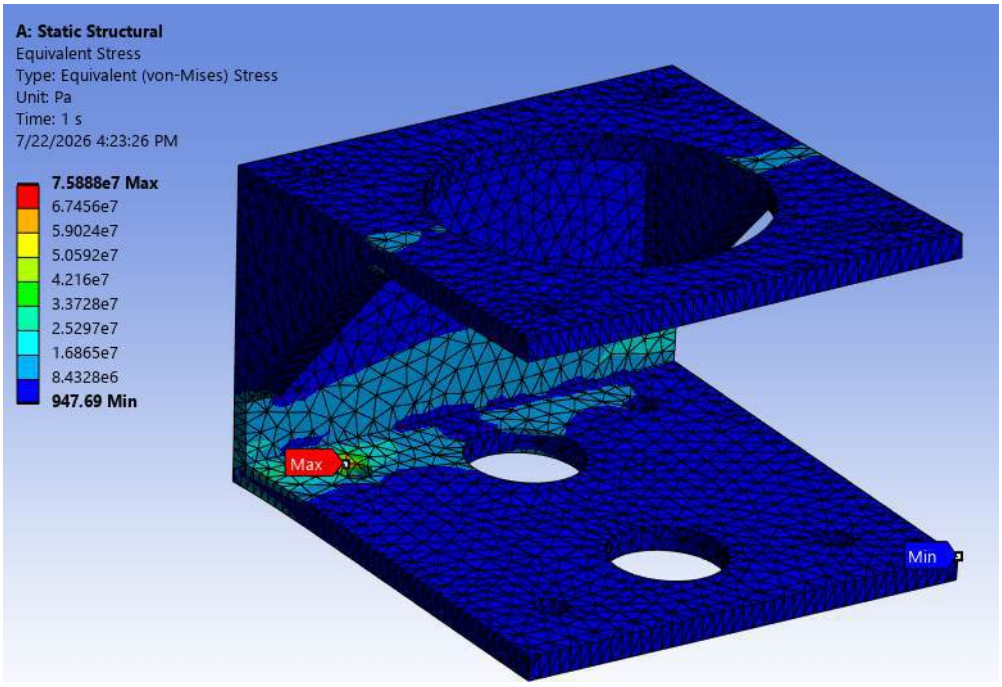


Fig. 3. ANSYS mesh and von Mises stress result (PA)

Python Active Learning Setup and Execution Loop

The active learning pipeline is executed entirely within a Python scripting environment, focusing on maximizing data efficiency via a pool-based uncertainty approach. From the master 200-point database, a permanent holdout verification set consisting of 40 design points is isolated to ensure objective downstream benchmarking. The Gaussian Process Regression model is then initialized using 10 random starting samples, which forms the seed training dataset (S_0). Ten points is a deliberately small starting budget, chosen to simulate the realistic situation a student researcher faces at the very beginning of a project, before any meaningful FEA data has been collected.

The optimization loop executes 40 sequential iterations according to the following algorithm: evaluate the GPR surrogate model on all remaining unlabeled members in the candidate pool; predict the mean stress (μ) and predictive standard deviation (σ) for each unlabeled candidate layout; apply the maximum uncertainty query criterion to isolate the single highest-uncertainty layout,

$$x^* = \text{argmax } \sigma(xi)$$

unmask the true high-fidelity FEA parameters for x* from the spreadsheet database, append this point to the active training set, and retrain the GPR model. Performance is measured at each individual query iteration against the permanent 40-point holdout set via root mean squared error (RMSE) tracking. The model convergence rate is directly benchmarked against a parallel control loop driven entirely by passive, random sample selection, using the same seed points and the same holdout set for both conditions.

Results and Discussion

The completed active learning loop converged on a candidate bracket geometry with a predicted mass of 27.04 g and a peak von Mises stress of 158.19 MPa, against a yield strength of 276 MPa for Al 6061-T6. Applying the margin of safety formula gives MoS = 276 / (158.19 × 1.5) − 1 ≈ 0.16, comfortably above the MoS ≥ 0 threshold and indicating the design would survive the 20g combined launch load case with a positive margin. Reaching this result required only 50 of the 200 total design points in the master database: 10 randomly seeded points plus 40 actively queried points, meaning three-quarters of the full simulation budget was never spent at all.

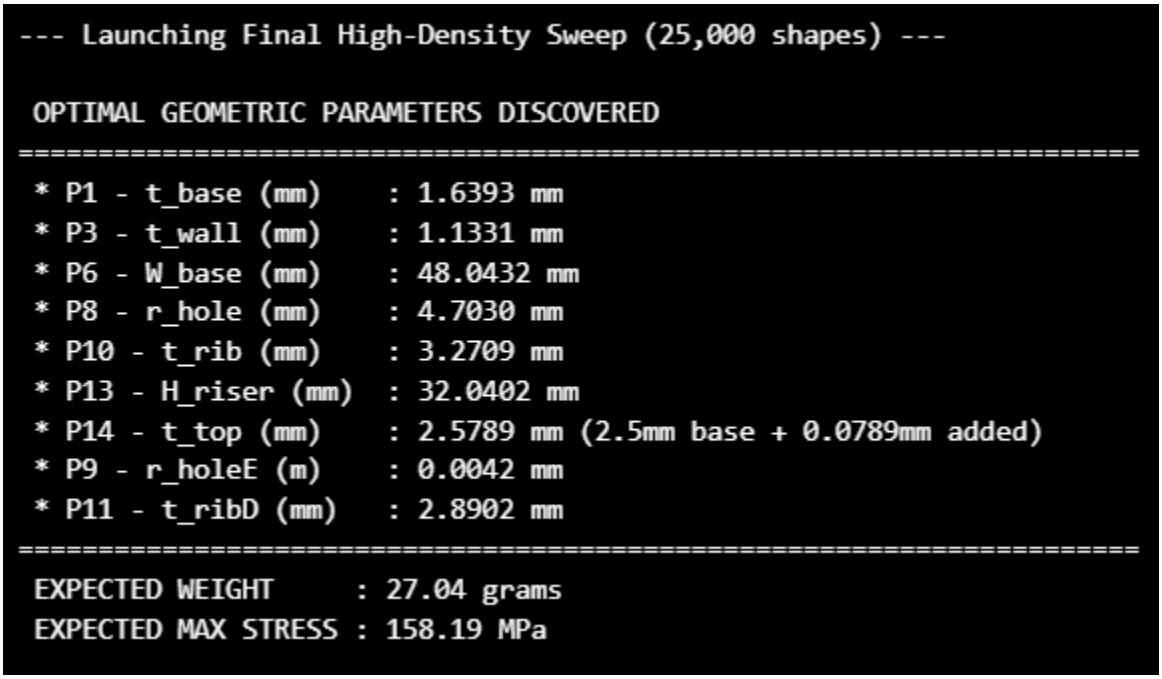


Fig. 4. optimal design summary output

That said, the optimizer's output is not fully clean. Two of the returned parameters , t_base (1.6393 mm) and t_wall (1.1331 mm) , fall below the operational bounds defined earlier (2.0–5.0 mm and 1.5–3.5 mm, respectively). Because the GPR surrogate was trained only on the interior of the sampled design space, any query resolving to an "optimum" outside that space is extrapolating past what the model actually learned, and extrapolated GPR predictions are known to be unreliable because the kernel's uncertainty estimate grows large and uninformative outside the training envelope (Rasmussen and Williams). Practically, this means the 27.04 g figure should be read as the model's best guess at a lower bound on achievable mass, not a manufacturable final geometry , a physically realizable version would need t_base and t_wall clamped back to their minimum bound values, pushing the true mass modestly above 27.04 g. This is best understood as a bounds-enforcement gap in the optimization script rather than a flaw in the active learning logic itself, and it is the first thing that should be fixed in any follow-up implementation of this pipeline.

The comparison between GPR and Random Forest as candidate surrogate families produced a clearer result than expected going into this study. Random Forest substantially outperformed GPR on holdout accuracy ($R^2 = 0.9672$, average error 4.36 MPa) compared to GPR's $R^2 = 0.7595$ and average error of 12.64 MPa; a linear Ridge Regression baseline landed between the two, at $R^2 = 0.8261$ and average error of 11.81 MPa. This is a meaningful departure from Ünlüsoy et al., whose benchmarking on a commercial aircraft wing surrogate found GPR and RF to be the two most competitive families with comparable trade-offs. The gap here is more likely explained by dataset scale , Ünlüsoy et al. had access to a training corpus roughly seven times larger than the 50-point actively-sampled set used in this study, and GPR's kernel-based interpolation is known to degrade faster than RF's tree-based partitioning when data is this sparse relative to the dimensionality of the design space. This is consistent with the broader pattern established earlier in the literature review: GPR and RF perform comparably in the abstract, but the specific dataset scale available to a given research team tips the balance toward one family or the other.

```
--- Leaderboard Results ---
[Linear Ridge Regression] R²: 0.8261 | Avg Error: 11.81 MPa
[Random Forest Regressor] R²: 0.9672 | Avg Error: 4.36 MPa
[Gaussian Process Regressor] R²: 0.7595 | Avg Error: 12.64 MPa
```

Fig. 5. R^2 and average error for Ridge, Random Forest, and Gaussian Process

The correlation matrix computed across the eight input parameters offers a possible structural explanation as well: r_hole and r_holeE, as well as t_rib and t_ribD, are each correlated at 1.00, as two pairs of the eight nominal input dimensions are not actually independent. This effectively collapses the real dimensionality of the design space below eight. In principle this redundancy should make the regression problem easier, not harder, but it also means the GPR kernel was fitting redundant, near-degenerate directions in input space , exactly the kind of pathology that can inflate predictive variance and depress R^2 without necessarily reflecting genuine model uncertainty about the underlying physics. A cleaner CAD parameterization that removed this redundancy before the Latin Hypercube sample was generated might have produced a better-conditioned GPR fit from the very first iteration.

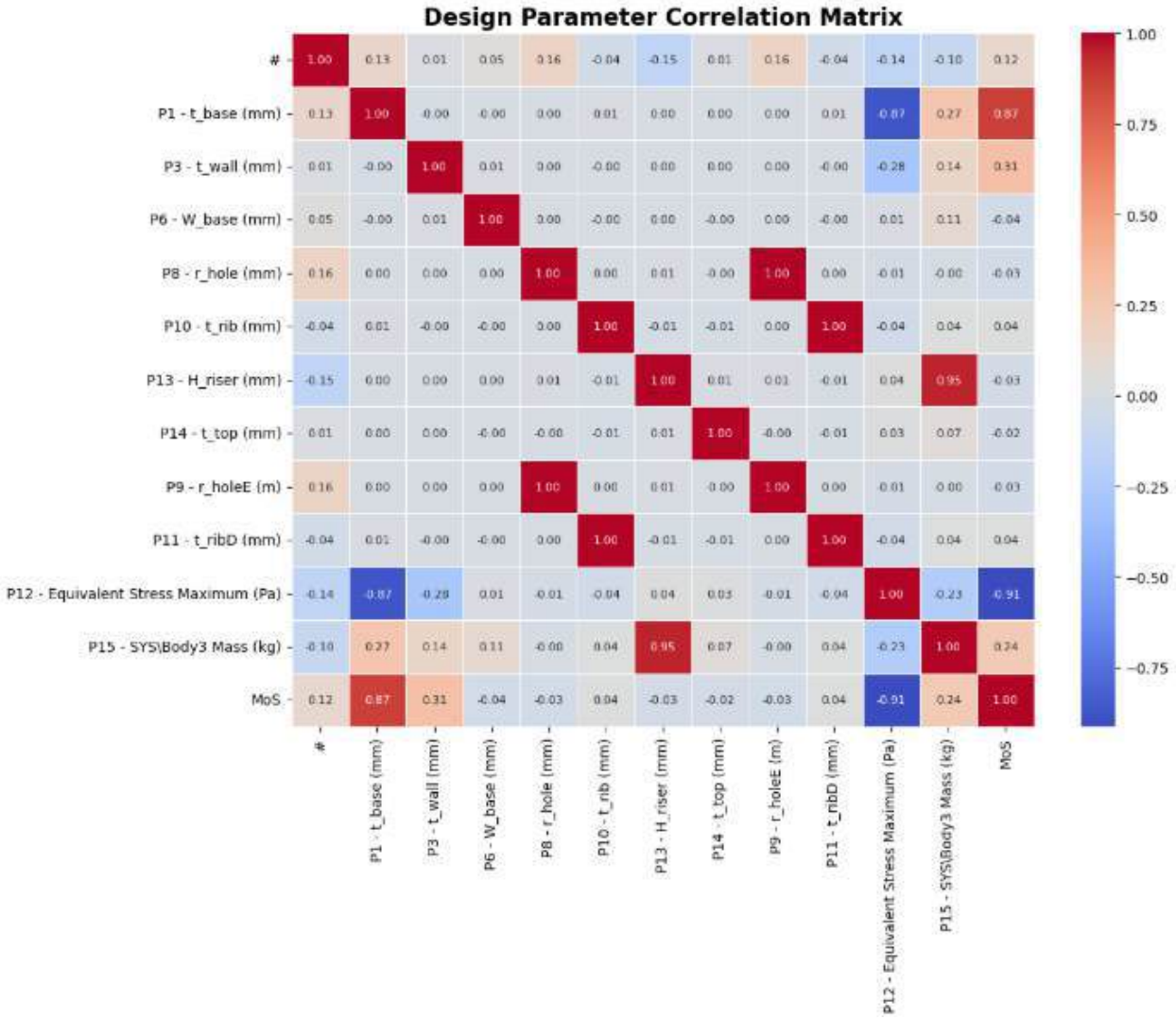


Fig. 6. design parameter correlation matrix heatmap

That degraded GPR fit is very likely the direct cause of the most consequential finding of this study: the active learning loop did not meaningfully outperform random sampling. Across all 40 queries, the AL-GPR RMSE curve and the random sampling RMSE curve tracked almost identically, both descending from roughly 86.55 MPa to roughly 86.20 MPa on the holdout set, with the two curves overlapping within their confidence bands for essentially the entire run. This stands in contrast to the significant reduction in solution time and data requirements that active learning is reported to provide over passive selection in a related decomposition-based optimization context , and it complicates the central premise of this paper's research question, which set out expecting a clear advantage for uncertainty-driven querying over passive sampling.

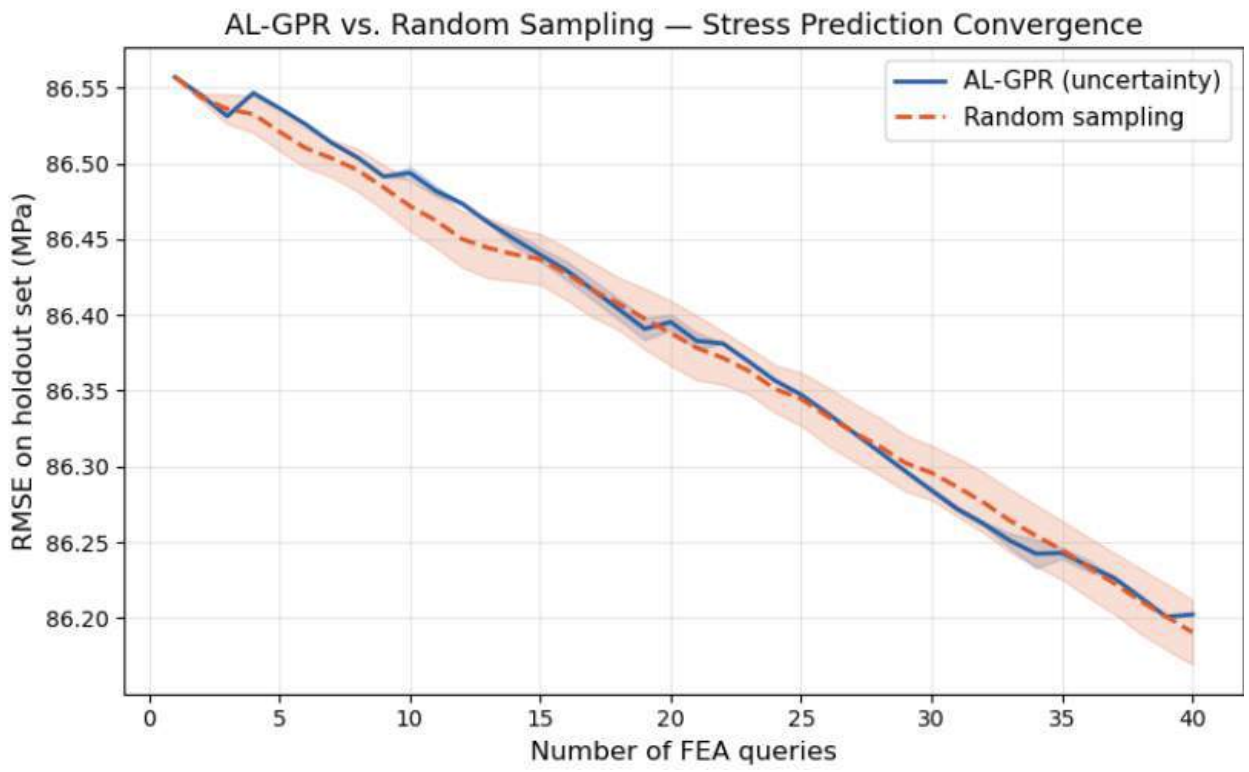


Fig. 7. AL-GPR vs. random sampling , RMSE convergence plot

The most likely explanation, given the R^2 disparity discussed above, is that the acquisition function $x^* = \text{argmax } \sigma(xi)$ is only as good as the uncertainty estimate producing it. If the GPR's $\sigma(x)$ values are not well-calibrated because the kernel is struggling with the correlated, sparsely-sampled eight-dimensional space, then the "highest uncertainty" point the loop selects at each iteration is not reliably the most informative point, and the loop degenerates toward behavior indistinguishable from random selection. Giovanis et al. note that adaptive strategies only outperform passive sampling when the exploration signal genuinely tracks regions of high information gain; the results here suggest that condition was not well met, likely due to the surrogate quality issue rather than a failure of the active learning framework itself. This is a different regime than the landing gear bracket study, whose own headline finding, that prediction accuracy for every surrogate family it tested improved consistently as dataset size grew from 500 to 5,000 samples, is in fact consistent with what this study observed at a much smaller scale: a surrogate family's practical usefulness cannot be assessed independent of how much data it was given. That study never tested a 50-point budget, but its trend line, extrapolated downward, predicts exactly the kind of degraded surrogate performance this study observed. These comparisons all imply the same underlying takeaway: GPR-driven active learning appears to require either a better-conditioned design space or a larger seed budget than was available in this study before its uncertainty signal becomes trustworthy enough to beat random sampling.

Conclusion

This study set out to test whether a pool-based active learning framework using GPR uncertainty could meaningfully reduce the number of high-fidelity FEA simulations needed to optimize a cislunar CubeSat star tracker bracket. The pipeline did succeed in producing a low-mass, structurally viable design, using only 50 of the 200 available design points, demonstrating that a small research team without cluster-scale compute can meaningfully explore an eight-dimensional bracket design space using nothing more than a spreadsheet-driven pool and a student ANSYS license. That part of the framework worked as intended and is directly reproducible by other independent researchers facing the same budget constraints described throughout the literature review. It also validates, independent of the active-versus-random comparison discussed below, that pool-based surrogate modeling is a viable design strategy for CubeSat-scale bracket hardware more broadly, regardless of which specific surrogate family or acquisition strategy a future researcher chooses to pair it with. Any student team working from an equivalent ANSYS student license and a comparably sized parametric design space should be able to reproduce a version of this workflow end to end, which was one of the explicit design goals stated at the outset of this paper.

Where this study diverges from its own hypothesis is in the comparison between active and random sampling: the two approaches converged to nearly identical RMSE over 40 queries, which does not replicate the large advantage reported for GP-based active learning in other engineering contexts. Rather than a failure of the framework, this is better read as evidence that the quality of the underlying surrogate model gates the value of uncertainty-driven acquisition , a Random Forest model fit the same 50-point dataset far more accurately, suggesting that a Random-Forest-driven acquisition strategy, or a hybrid approach retaining GPR purely for its native uncertainty quantification, may be a more effective architecture for this class of problem than a GPR-only pipeline, particularly for research teams who, like this one, are operating with a training budget in the tens rather than hundreds of points.

Limitations and Future Scope

Two limitations should guide follow-up work. First, the optimizer's final geometry included two parameters (t_base, t_wall) outside their defined operational bounds, indicating the optimization script needs explicit bound-clamping before its output can be treated as a manufacturable design rather than a theoretical lower bound. Second, this study operated on a single 200-point Latin Hypercube pool and a single material, Al 6061-T6; a natural next step, beyond scaling this no-code framework to entire multi-component CubeSat chassis assemblies, would be re-running the same comparison on a larger seed pool to test whether the near-identical convergence observed here is a permanent property of this design space or an artifact of GPR's data requirements at this scale. A third avenue worth pursuing is re-parameterizing the CAD model to remove the two fully correlated dimension pairs identified in the correlation matrix, which would give the GPR kernel a cleaner, lower-dimensional space to work with and might restore the advantage active learning is expected to provide.

Ultimately, this work still supports the broader case made throughout this paper: data-efficient, no-code machine learning pipelines can put real structural optimization capability into the hands of student and independent spaceflight teams who will never have access to a supercomputer. But the specific choice of surrogate model within that pipeline matters more than this study initially assumed, and that choice deserves the same rigor currently reserved for the FEA and CAD stages of the workflow. For a field where every gram and every hour of compute time carries real mission consequences, that is not a small correction to make. A student researcher working through this same problem in the future would be well served by benchmarking surrogate families before committing to an acquisition strategy, rather than assuming, as this study initially did, that GPR's theoretical elegance would translate directly into practical accuracy at a severely constrained data budget.

Future researchers attempting a similar pipeline on a different CubeSat component should treat the benchmarking step, comparing candidate surrogate families against each other before committing to one as the acquisition engine, as a required part of the methodology rather than an optional check, since this study's results suggest that step is where the real risk to a successful active learning implementation lies.

The effective answer this study arrived at is conditional: active learning reduces the simulation budget dramatically relative to exhaustive search under any surrogate, but it only outperforms passive random sampling only when the underlying surrogate's uncertainty estimates are well-calibrated, which was not reliably the case for GPR at this study's data scale.

Works Cited

Barsoum, A., et al. "Structural Analysis of 1U CubeSat Designed for Low Earth Orbit Missions." ResearchGate, 2023, www.researchgate.net/publication/374412233.

"Comparative Analysis of Surrogate Models for Nonlinear Behavior Prediction of an Aircraft Landing Gear Bracket." ScienceDirect, 2026, www.sciencedirect.com/science/article/pii/S2307187726000040.

"CubeSat Design Specification (Rev. 13)." The CubeSat Program, Cal Poly SLO, 2014, static1.squarespace.com/static/5418c831e4b0fa4ecac1bacd/t/56e9b62337013b6c063a655a/1458157095454/cds_rev13_final2.pdf.

"Data Efficient Surrogate Modeling for Engineering Design: Ensemble-Free Batch Mode Deep Active Learning for Regression." arXiv, 2022, arxiv.org/abs/2211.10360.

Giovanis, D. G., et al. "Active Learning for Adaptive Surrogate Model Improvement in High-Dimensional Problems." Structural and Multidisciplinary Optimization, vol. 67, no. 7, 2024, article 122, doi.org/10.1007/s00158-024-03816-9.

Hoffer, J. G., et al. "Gaussian Process Surrogates for Modeling Uncertainties in a Use Case of Forging Superalloys." Applied Sciences, vol. 12, no. 3, 2022, article 1089, doi.org/10.3390/app12031089.

Ma, Z., et al. "Attitude Correlated Frames Based Calibration Method for Star Sensors." Sensors, vol. 24, no. 1, 2024, article 67, pmc.ncbi.nlm.nih.gov/articles/PMC10781283/.

Mecanica Solutions. "Aerospace CAD Prototyping." Mecanica Solutions, mecanicasolutions.com/en/blog/aerospace-cad-prototyping/. Accessed July 2026.

Rasmussen, Carl Edward, and Christopher K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006, gaussianprocess.org/gpml/.

Rocket Lab. "Rocket Lab Launches New Constellation-Class Star Tracker." Rocket Lab, 13 Apr. 2023, www.rocketlabcorp.com/updates/rocket-lab-launches-new-constellation-class-star-tracker

Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009, burrsettles.com/pub/settles.activelearning.pdf.

Sauer, Annie, et al. "Active Learning for Deep Gaussian Process Surrogates." Technometrics, vol. 65, no. 1, 2023, pp. 4-18, doi.org/10.1080/00401706.2021.2008505.

"Actively Learning Deep Gaussian Process Models for Failure Contour and Probability Estimation." AIAA SciTech Forum, 2024, doi.org/10.2514/6.2024-0577.

Shao, Y., et al. "Accelerating Analysis for Structure Design via Deep Learning Surrogate Models." Advanced Intelligent Systems, 2023, doi.org/10.1002/aisy.202200099.

Shields, B. J., et al. "Bayesian Reaction Optimization as a Tool for Chemical Synthesis." npj Computational Materials, vol. 7, 2021, article 100, doi.org/10.1038/s41524-021-00656-9.

Sravanthi, K., et al. "Design and Analysis on 1U CubeSat Using Different Materials." NeuroQuantology, vol. 20, no. 16, 2022, pp. 4942–4954, doi.org/10.14704/nq.2022.20.16.NQ88491.

"Taking the Human Out of Decomposition-Based Optimization via AI, Part II: Learning to Initialize." arXiv, 2023, arxiv.org/abs/2310.07082.

Ünlüsoy, C., et al. "Analysis of Wing Structures via Machine Learning-Based Surrogate Models." Aerospace, vol. 13, no. 4, 2026, article 338, doi.org/10.3390/aerospace13040338.

Applications of Artificial Intelligence in Eczema Management: A Scientific Literature Review By Aastha Chivakula

Abstract

Eczema is a chronic inflammatory skin disease that affects millions of people worldwide and often requires long-term management. Recent advances in artificial intelligence (AI) have led to new tools that support different aspects of eczema care. This literature review examines research published between 2020 and 2026 on the use of AI for eczema diagnosis, severity assessment, symptom monitoring, and treatment planning. Current studies show that AI models can accurately identify eczema from clinical images, estimate disease severity using standardized scoring systems, monitor symptoms such as itching through wearable devices and mobile applications, and assist clinicians in selecting personalized treatments. However, several challenges remain, including limited representation of diverse skin tones in training datasets, algorithmic bias, differences in evaluation methods, and a lack of long-term clinical validation. Overall, the evidence suggests that AI has strong potential to improve eczema care, although further research is needed before these technologies can be widely adopted in clinical practice.

Keywords: Artificial Intelligence; Atopic Dermatitis; Dermatology; Eczema; Machine Learning; Precision Medicine

Section 1 Introduction

Eczema is a type of dermatitis that causes your skin to become dry, pruritic and inflamed. Eczema is associated with dysfunction of the skin barrier, reducing its ability to retain moisture and protect against environmental irritants.[1] In the US alone, approximately 10% of the population (about 31.6 million people) struggle with eczema. African American individuals are more likely to have eczema and their conditions tend to be more severe. In addition, 80% of individuals who develop eczema start showing symptoms before age 6, showing that children are also heavily affected.[2]

Artificial Intelligence (AI) has emerged as a growing tool for eczema diagnosis and treatment; however, problems continue to persist. By analyzing clinical datasets, medical images and patient reports, AI systems may assist clinicians in diagnosis, monitoring and treatment planning. AI models can be updated as new data become available, allowing them to identify complex patterns that may support medical diagnosis and treatment decisions.[3]

Traditionally, the diagnosis of eczema is based on clinical evaluations of skin lesions, medical history reviews and allergy or patch tests. Diagnosis is heavily based on visual observation and in extreme cases, invasive diagnostic procedures are necessary, such as biopsies.[4] Additionally, treatment typically focuses on restoring the skin barrier, and minimizing inflammation through approaches such as moisturization, topical medications, phototherapy and systemic therapies.[5] Because eczema shares clinical features with several inflammatory skin disorders, accurate diagnosis remains challenging. Visual assessment can be subjective and may

contribute to diagnostic variability, creating interest in AI-based image classification systems.[6] Due to substantial differences in disease presentation, creating treatment plans and long term management can be difficult. Image classification models are among the most common AI approaches used in eczema research.[7] These systems are trained on large collections of clinical images to recognize patterns associated with eczema and distinguish affected skin from healthy skin.[8]

The growing body of research on AI in eczema care suggests that AI-driven systems may support more consistent, accessible, and personalized disease management across multiple stages of care. Studies have demonstrated applications in diagnosis, severity assessment, symptom monitoring, and treatment planning, highlighting the potential of AI to support clinical decision-making and patient management. Although many of these approaches have shown promising results in research settings, concerns remain regarding dataset quality, population diversity, model generalizability and clinical implementation. This literature review evaluates the current evidence for AI in eczema care, focusing on its role in diagnosis, severity assessment, symptom monitoring, and treatment planning, as well as the limitations and future challenges facing these technologies.

Methodology

This literature review was conducted to evaluate the current applications of AI in eczema diagnosis, severity assessment, symptom monitoring, and treatment planning. Relevant sources were identified through searches of scientific databases and academic search engines, including PubMed, Google Scholar, JMIR, arXiv, and peer-reviewed dermatology journals. Search terms included combinations of keywords such as "eczema," "atopic dermatitis," "artificial intelligence," "machine learning," "deep learning," "diagnosis," "severity assessment," "symptom monitoring," and "treatment prediction."

Studies published primarily between 2020 and 2026 were prioritized to capture recent developments in AI technologies. Articles were included if they examined AI-based approaches related to eczema care and reported information regarding model design, datasets, performance metrics, or clinical applications. Review articles, primary research studies, conference proceedings, and selected emerging research publications were considered when they provided relevant information about the topic. Sources unrelated to eczema or lacking sufficient methodological detail were excluded.

The selected studies were organized into four major categories: diagnosing eczema, assessing disease severity, monitoring symptoms, and determining treatment plans. Key findings, datasets, AI methodologies, performance measures, and limitations were analyzed and compared to identify major trends, strengths, challenges, and future research directions in the use of AI for eczema management.

Section 2 Main Review Section

(2.1) Diagnosing Eczema

Delayed or inaccurate diagnosis of eczema can lead to reduced quality of life, as patients can experience painful rashes, inflammation and in some cases infections all before diagnosis. Early diagnosis and targeted treatment plans can help mitigate these effects and improve patients' quality of life. Traditionally, dermatologists diagnose eczema through a series of clinical evaluations based on patient history, appearance of skin and physical exams. If more information is needed, dermatologists turn to skin biopsies and other methods of invasive testing.

Around 2020, researchers began training AI models to assist in the diagnosis of eczema. However, due to the lack of available datasets at the time, the amount of research on the subject was relatively small. The fast rise of deep learning in 2021 and 2022 has allowed medical imaging to become much more prevalent. Researchers Rasheed et al. developed a hybrid AI model that combined deep learning image analysis with a ML classifier to improve diagnostic accuracy. This work contributed to the creation of a multiclass image dataset: the Eczema Image Resource (EIR), a dataset containing 2,039 labeled clinical images, specifically designed to train AI systems. The model was tested by comparing its performance with other ML and deep learning models using evaluation metrics like accuracy, sensitivity, and specificity. The Hybrid-6 model achieved the highest performance, and was selected by the researchers. It reached an accuracy of 88.29% and specificity of 90.33%. [9] These findings suggest that AI based image analysis may help improve diagnostic consistency. However, this study was evaluated under controlled conditions, raising questions about how well the model would generalize to more diverse patient populations and real world settings. Despite this, the study represented an important early step in the development of AI assisted eczema diagnosis by providing a large dataset and framework for future research.

Many individuals with eczema may not receive a formal diagnosis. A population-based survey in the US found that among individuals reporting eczema symptoms only one third (37.1%) of cases reported having a physician confirmed diagnosis.[10] Although AI systems can analyze large datasets, accurate diagnosis requires quality, diversity and sufficient clinical relevance of the data used for training. Researchers Kalbande et al., aimed to address this by designing an affordable, easy to use, smartphone application that allows users to enter a photo of their skin, and receive a probability estimate of whether they have eczema. The system used image processing techniques and convolution neural networks (CNN) to identify visual patterns associated with eczema from images collected by medical professionals. When evaluated on a dataset using 50 images of eczema-affected skin and 50 images of normal skin, the model was able to identify 256 features from each image and match them to attributes of eczema. The system had approximately 85% accuracy in detecting eczema and providing probability estimates. [11] This smartphone application demonstrated potential as an accessible approach for preliminary eczema screening. However, this study was limited by its small dataset and the fact that these systems rely heavily on visual input. Factors such as lighting, quality, and differences in skin characteristics could influence model performance, highlighting a need for larger datasets before clinical implementation.

Purely based on visual data, eczema may be complicated to diagnose because other diseases present similar symptoms. Ding et al. developed a web-based tool focused on improving the differential diagnosis of eczema versus psoriasis. Eczema and psoriasis often have overlapping symptoms (e.g. inflamed, red skin) which makes proper diagnosis based on appearance alone challenging. Researchers in China developed an ML-based tool that uses laboratory data to provide a faster, noninvasive alternative. Instead of relying only on images, the researchers incorporated clinical and laboratory data from more than 57,518 adult patients with eczema or psoriasis and examined biological differences, finding 16 variables that significantly differed between the two diseases. Researchers compared multiple ML approaches and developed a combined model designed to improve robustness and accuracy. This model, the XGBoost (Extreme Gradient Boosting) model, achieved an area under the curve (AUC) of 0.812, where AUC measures how well a model distinguishes between conditions. Researchers also applied SHAP (Shapley Additive Explanations), an explainable AI method, to improve interpretability by showing how individual features influenced predictions. This resulted in a user-friendly web-based diagnostic system with an accuracy of 73.13% [12], allowing clinicians to either input lab values or upload images and receive real-time probability predictions. This study highlights how multi-model clinical data improves differential diagnosis, as it often requires more than visual or surface-level data because underlying biological differences between skin conditions can also influence diagnostic accuracy.

Furthering biological insights can continue to improve diagnostic understanding. Eczema is not only similar to other skin diseases, but has also been found to co-occur with asthma and food allergies, as these conditions may share genetic or immune pathways. Researchers created PandaOmics, an AI-driven target discovery platform made to identify common biological targets. It uses AI to analyze large amounts of genetic data to find potential targets shared between all three conditions. The study included 32 total case-control comparisons. The researchers examined transcriptomics data, showing which genes are turned on or off in patients with these diseases. PandaOmics prioritized the top 100 targets for further analysis, finding six high-confidence targets that were shared across all three diseases.[13] These shared targets may provide insight into common disease pathways and highlight potential areas for future therapeutic research. This may help guide future drug development and improve outcomes for patients who suffer from more than one condition simultaneously.

Biologically eczema may not be an isolated condition, but rather part of a broader group of immune-related diseases that complicate both diagnosis and treatment. This inclusion of non-image data in more recent models highlights a shift in the literature toward multimodal diagnosis, suggesting that visual analysis alone is not enough to capture the full biological complexity of eczema and related skin conditions. PandaOmics demonstrates how AI can move beyond visual assessment towards identifying molecular and genetic patterns, which may improve diagnosis accuracy.

In addition to biological complexity, another major limitation in current diagnostic AI systems is the lack of diversity in training datasets, which can affect accuracy across different

patient populations. Many existing studies using AI for eczema diagnosis use datasets that have limited representation of darker skin tones because many of the test subjects are fair skinned. In darker skin, common signs of eczema such as redness can be more difficult to see due to skin pigmentation, and since many of the diagnostic criteria that doctors use to diagnose eczema were developed based on lighter skin types, accuracy may decrease when tested on darker skin tones. A study conducted in Nigeria aimed to address this difficulty. Researchers developed and tested an ML algorithm designed to detect eczema and distinguish it from healthy skin in individuals with darker skin types. The study collected clinical images from patients, all with darker skin, at the Dermatology Clinics of Obafemi Awolowo University Teaching Hospitals Complex. The dataset was designed to be diverse, including a variety of eczema severity levels, body locations, sexes, and age groups, to improve model generalizability. Researchers used an EfficientNet-based CNN, training it based on 1,403 labeled images of eczema and healthy skin. To test the model's performance, 255 additional images were used. The algorithm reported strong performance, reaching approximately 91% accuracy while correctly identifying most eczema lesions across darker skin tones [14], indicating that dataset diversity plays a critical role in model performance and that AI systems trained on representative populations can achieve strong diagnostic accuracy in darker skin types. Since research with darker skin tones is limited, further research is still needed to determine whether these models maintain performance in broader clinical settings.

Collectively, these studies demonstrate the evolution of AI-based eczema diagnosis from simple image classification models to more advanced systems that integrate clinical data, genetic information, and diverse patient populations. Although diagnostic accuracy has improved, further research is needed to ensure these models remain reliable and effective across different patient groups and real-world clinical settings.

Study	AI Method	Dataset	Key Results
Hybrid-6 Model (2022)	Hybrid deep learning + ML	2,039 eczema images	88.29% accuracy; 90.33% specificity
Kalbande et al.	CNN smartphone app	100 skin images	~85% accuracy
Eczema vs. Psoriasis Study	XGBoost + SHAP	57,518 patients	AUC = 0.812; accuracy = 73.13%
PandaOmics	AI genomic analysis	32 case-control comparisons	Identified 6 shared biological targets
Nigeria Study	EfficientNet CNN	1,403 training images; 255 test images	~91% accuracy

(2.2) Assessing Severity of Eczema

Eczema severity can range from mild (small, infrequent dry patches) to severe (widespread, intense itching, cracking and bleeding) and treatment options can differ based on the severity. Assessing the severity of eczema is essential for determining proper treatment plans to help reduce harsh symptoms. Accurate severity assessment also helps clinicians monitor disease progression and evaluate treatment effectiveness.

One difficulty in developing appropriate treatment plans for eczema is accurately determining the severity of a patient's skin lesions. The appearance of the disease on the skin does not always match the severity of patients' reported symptoms, particularly itching. A 2025 study published in AJMC investigated whether AI could support severity classification using patient photographs. The system consisted of three components: body part detection, lesion identification, and severity evaluation using the Three-Item Severity (TIS) score, which rates redness, swelling, and skin damage on a scale from 0 to 9. Researchers trained the model using 880 labeled images from the Atopiyo platform, a website where patients upload photos of their skin conditions. The model was then evaluated using 9,656 images from 900 participants. The model identified body parts with approximately 98% accuracy and detected eczema lesions with 100% accuracy. When compared with evaluations from dermatologists and allergists using 220 images, the results showed strong agreement ($R = 0.73$, $P < .001$). However, the correlation between AI severity scores and patient-reported itching levels was weak ($R = 0.11$), indicating a gap between clinical severity and perceived symptoms.[15] These findings suggest that visual severity and patient symptoms may not always align, highlighting a challenge in assessing eczema severity using photographs alone.

Beyond static assessment, researchers have also studied how eczema severity changes over time, as symptoms can fluctuate unpredictably. Periods of worsening symptoms, known as flares, may be followed by improvement, making it difficult for clinicians to determine the most appropriate treatment at a given time. Researcher Hurault et al. developed a ML model designed to predict daily severity fluctuations in individual patients, which in turn may help doctors to time treatments better to manage the disease. Researchers trained the model using data from 59 children who were followed for six months and 334 children followed for 16 weeks. The model applied Bayesian inference to update predictions as new data became available. To validate, researchers used forward-chaining validation to simulate real-world prediction using only prior observation, meaning the models predicted future severity based only on past patient data, similar to how it would work in real life. The results showed that the model could successfully predict daily eczema severity at the individual level and improved prediction accuracy by about 60%, while also identifying patient-specific disease patterns, including symptom duration and treatment response variability. [16] These findings suggest that machine learning approaches may help track eczema progression over time in a clinically relevant manner, but further research needs to be conducted on a larger, more diverse population.

Another major challenge in managing eczema is that many patients, especially children, do not consistently follow their treatment plans, which can make the condition worse over time. Researchers created a mobile health app, called Atopic App, to help families better monitor and

manage eczema in children. It allows caregivers to upload photos of a child's skin, track symptoms, and receive treatment guidance. The AI model was designed to estimate eczema severity using the Eczema Area and Severity Index (EASI), a medical scoring system doctors use to measure how much skin is affected and the severity of the redness, swelling, and irritation. To train the model, dermatologists collected and labeled images of eczema and marked the affected skin areas so the AI model could learn to recognize patterns of the disease. The study analyzed data from 1,223 app users, including 910 caregivers, who uploaded a total of 3,385 pictures of children with eczema. Researchers also compared the model's estimated EASI scores with the Patient-Oriented Eczema Measure (POEM), a questionnaire that measures symptoms such as itching, sleep loss, and skin irritation. The results showed a moderate correlation ($R = 0.58$, $P < .001$); when the AI predicted more severe eczema, patients usually reported worse symptoms as well. The study also evaluated the feasibility, or practicality, of using the app in everyday life and found that many families were able to regularly upload photos and track symptoms, showing that the technology could realistically be used at home. In addition, among users who completed follow-up surveys about three to four weeks later, about 70.6% reported lower eczema severity scores. This suggests that AI-based severity tracking may help support families, improve accessibility and increase patient engagement.[17] However, model accuracy remains dependent on the quality and consistency of user-generated data. Follow-up data was limited to approximately three to four weeks, and some outcomes were based on self-reported measures, which may introduce variability.

EczemaNet is a similar automated remote assessment tool that allows users to input a photo of eczema-affected skin. The app uses image classification to evaluate the severity of the eczema, in accordance with standard clinical scoring systems. EczemaNet focused on making severity assessment more objective and consistent for children and adolescents. All of the images used for training were taken from individuals under age 18 years who had been diagnosed with eczema. Researchers took pictures of four body regions: the head and neck, body, arms, and legs. Each image was graded on a scale from 0 to 3, using in-person medical assessments as the reference. After being trained on 920 images from 52 patients, it measured what percent of the skin was affected and then calculated an overall severity score for each patient. To evaluate performance, it was tested on 2,342 images, collected from 252 patients. The results were compared to standard clinical scoring systems, EASI and validated Investigator's Global Assessment (vIGA), a 5 point (0-4) clinician scale to measure overall severity. The approaches that were designed to mimic EASI showed correlation scores of about 0.92 to 0.93, while those mimicking vIGA had correlations between 0.82 and 0.92. The study also measured agreement between the AI and doctors, which showed strong agreement, with scores ranging from about 0.71 to 0.73. [18]

While EczemaNet showed benefits, it was not fully automatic as it still needed people to tell the system where the eczema was before it could assess severity. In response, the AI system EczemaNet2 was developed to automatically determine the severity of eczema from standard digital photographs, which may not fully replace the need for in-person clinical evaluations, but

could possibly reduce the frequency of visits. This may be beneficial as in-person evaluations can be time consuming and are prone to variation. EczemaNet2 was trained using 1,345 clinical images from 287 children aged 6 months to 16 years with eczema, dermatologists manually labeled the eczema regions at the pixel level, allowing EczemaNet2 to learn exactly where eczema appears. The dataset also included a range of skin tones, helping improve the ability to work across diverse patients. To test and validate, researchers evaluated EczemaNet2 on 271 images for lesion detection and 136 images for severity prediction. EczemaNet2 achieved an average precision and F1 score of about 60% for detecting eczema areas and improved detection quality by approximately 25% compared to the earlier EczemaNet system. It maintained similar accuracy in predicting severity scores.[19] Unlike previous approaches, this study uses detailed dermatologist-labeled pixel annotations and real-world photographs with variations in lighting, blur, and background conditions, making the system more practical for remote monitoring. Because the model was evaluated using real-world photographs with natural variations, the findings suggest that EczemaNet2 may have greater potential for clinical and remote monitoring applications. EczemaNet2 represents one of the more advanced automated systems developed for eczema severity assessment.

AI severity-assessment models can provide consistent evaluations that closely match clinician ratings, improving the objectivity of eczema severity measurements. However, differences in scoring methods and gaps between clinical assessments and patient-reported symptoms highlight the need for further research to develop more comprehensive systems.

Study	AI Method	Dataset	Key Results
Japan Severity Study	AI lesion detection and TIS scoring	9,656 images from 900 participants	98% body part detection; strong agreement with specialists (R = 0.73)
Daily Severity Prediction	Bayesian ML model	393 children	Predicted daily severity changes and improved accuracy by ~60%
Atopic App	AI-based EASI scoring	1,223 users; 3,385 images	Moderate correlation with symptoms (R = 0.58); 70.6% reported improvement
EczemaNet	Automated image analysis	920 training images; 2,342 test images	Strong agreement with EASI and vIGA severity scores
EczemaNet2	Automated lesion detection and severity assessment	1,345 images from 287 children	Improved lesion detection by ~25% while maintaining severity accuracy

(2.3) Monitoring Symptoms

Eczema usually manifests through intense itching, dry and sensitive skin, and recurring rashes that can significantly reduce an individual's quality of life by affecting sleep cycles and causing persistent physical discomfort. Continuous monitoring of symptom progression is essential for developing targeted treatment plans because symptoms often fluctuate over time. By identifying patterns in symptom fluctuation, doctors can assess what treatment methods are effective and predict future flare-ups, making symptom monitoring an important area for the application of AI.

Bibi et al. proposed AutoAI, a smartphone-based monitoring system that allows users to track their eczema symptoms and predict disease progression. AutoAI analyzed itching levels, pain, and quality-of-life scores using a combined dataset from three clinical studies involving 368 patients with 67 different features who had chronic hand or foot eczema. Participants were required to collect symptoms over a period of about six months in these studies, allowing the model to identify patterns in disease progression. The app then analyzed patterns in these symptoms and tried to predict how they might change in the future. Researchers compared multiple ML approaches and selected the best-performing model for each symptom prediction task. AutoAI demonstrated high predictive performance for itching patterns ($r^2 = 0.90$), while predictions for pain and quality-of-life outcomes were slightly less accurate, with performance in the low 0.80 range. These findings may indicate that AI-based symptom monitoring is most effective for measurable physical symptoms like itching, while more subjective experiences such as quality of life are harder to predict.[20]AI-based monitoring systems may also be used by clinicians to help them learn how a specific patient's symptoms change over time so they can provide better patient care, highlighting the potential of AI to support long-term symptom monitoring and provide clinicians with additional information about disease progression between office visits.

While app-based systems rely on patient-reported data, newer approaches use sensor-based ML models to objectively detect symptoms such as scratching and sleep disruption. Chronic itch, one of the main symptoms of eczema, often leads to serious sleep disruption and reduced quality of life. Doctors and researchers do not currently have a reliable and objective way to measure chronic itch and how it affects sleep, since most clinical evaluations depend on patients self-reporting their symptoms, which can be inconsistent or inaccurate. To address this challenge, researchers Ouroutzoglou et al. monitored 20 adults diagnosed with chronic itch conditions, including eczema, in their homes for four weeks. Data was collected using a radio frequency sensing device, which sends low power radio signals into the room and analyzes how the signals change when the person moves, along with an infrared video camera to record events, maintaining accuracy. The goal was to detect scratching behavior and quantify sleep quality without requiring wearable devices, which can be uncomfortable for patients with sensitive skin. The dataset included 364 total nights of recordings and 21,619 labeled scratching events. The model achieved an area under the curve (AUC) of 0.997 for detecting scratching events longer than three seconds, along with a specificity of 99.7 percent and a sensitivity of 82.5 percent. For

the most part, it rarely mislabeled normal movement as scratching and correctly identified most real scratching events.[21] The study also found a correlation between scratching and worse sleep outcomes, such as reduced sleep efficiency and longer time needed to fall asleep, highlighting the value of objective symptom monitoring for guiding future treatment strategies.

Nocturnal scratching is another one of the most common symptoms of eczema that significantly impair an individual's quality of life, which also relies on patient reported data. Researchers Xing et al. developed two kinds of ML models for scratch prediction using wrist wearing actigraphy data. The first model used physical motion features to train a gradient boosting classifier, while the second model used a deep learning approach with CNNs to predict scratching directly from movement signals. Initial data was collected from seven patients. Participants wore wrist-based sensors during sleep while synchronized video recordings were used to label scratching events. Researchers trained models to distinguish scratching from other movements and evaluated performance on unseen participants. Based on the data collected from the watches and video tracked motions researchers concluded 13,335 segments were associated with non scratch activity and 297 segments were linked to scratch activities. Due to this imbalance, the researchers used training techniques to help the model focus more on scratch events rather than being overwhelmed by non-scratch data, further dividing the segments into train, validate, and test groups. Results demonstrated that the model achieved an F1 score of 0.45 on the test set, accompanied by a precision of 0.44.[22] Although the study was limited by its small sample size and potential difficulty capturing subtle finger-only scratching motions, the researchers concluded that the automatic scratch detection algorithm holds the potential for objectively assessing sleep quality and disease state in eczema patients. This can further help to inform and refine strategies to improve quality of life for individuals with eczema.

Since nocturnal scratching has traditionally been treated through medication due to lack of tools, Yang et al. monitored 10 adult (six female, four male) patients with mild eczema who reported moderate to severe nighttime scratching and monitored them in their own homes using a wearable sensor. The sensor utilizes AI to detect scratching and a haptic feedback feature that causes the device to vibrate when scratching is detected to interrupt the scratching behavior. First, each patient wore the device for seven nights with only the sensing feature activated, to collect baseline data on their nocturnal scratching patterns. Then, for the next seven nights, each patient wore the device with the haptic feedback feature activated, so that when the AI model identified scratching, the device would gently vibrate, notifying the patient of their itching behavior so that they can consciously stop itching. The device used an AI model trained to recognize scratching motion patterns from wearable sensor data. To test how well this system worked, the researchers compared the number of scratch events each night and the total duration of scratching per hour of sleep before and after activating the haptic feedback. The results showed a significant reduction in scratching when the vibration feedback was activated, with the average total number of nightly scratch events falling by 28 percent, and the average scratch duration per hour of sleep cut by 50 percent.[23] This wearable system may serve as a useful non-clinical method to reduce nighttime scratching in people with eczema, but larger and more

controlled studies are needed to confirm these findings.

Beyond monitoring current symptoms, researchers have also explored whether symptom patterns can be used to predict future disease progression. Monitoring symptoms for eczema is difficult because patients often report fluctuating symptoms, with the most severe symptoms being classified as “flare ups”. Current clinical methods do not reliably consider these flare ups when assessing the disease or planning treatments, which can reduce the effectiveness of the treatments. Researchers Nielsen et al. used data consisting of 878 patients from the Danish Skin Cohort, to analyze how past flares can predict future disease severity. The research team studied how the number of flare ups reported in 2022 was associated with multiple patient-reported severity measures collected in 2023. Most patients experienced flare ups, with 405 reporting between one and five flare ups per year and 169 reporting six to ten flare ups, and researchers concluded that patients with higher flare frequency, especially those with more than 10 flares per year, had significantly higher severity and lower quality of life, making flare rate a strong predictor of eczema outcomes.[24] Researchers concluded that AI models can provide objective and accurate predictions of disease courses in eczema, meaning that tracking flare up symptoms can help clinicians better understand which patients are at higher risk and allow for more informed planning and patient management.

Overall, AI-based symptom monitoring has expanded beyond patient self-reporting by incorporating smartphone applications, wearable sensors, and contactless monitoring technologies. Although many systems still require larger validation studies, current findings suggest that AI may improve long-term symptom monitoring and enable earlier intervention as eczema symptoms worsen

Study	AI Method	Dataset	Key Results
AutoAI App	Machine learning symptom prediction	368 patients	Predicted itching with high accuracy (R^2 = 0.90) and pain/quality of life in the low 0.80 range
Chronic Itch Sleep Monitoring (2025)	Radio-frequency sensing + ML	20 adults; 364 nights	AUC = 0.997; detected scratching with high sensitivity and specificity
Nocturnal Scratch Detection (2024)	CNN and Gradient Boosting	7 patients; 13,632 movement segments	F1-score = 0.45; demonstrated potential for automated scratch detection
Wearable Haptic Feedback Device	AI scratch detection with vibration feedback	10 adults	Reduced scratch events by 28% and scratch duration by 50%

(2.4) Determining Treatment Plans

Effective eczema management depends on selecting treatments that are appropriate for an individual patient's disease severity, symptoms, and risk factors. Because eczema is highly heterogeneous, developing personalized treatment plans remains a major challenge in dermatology.

Genetics, immune system issues, environmental triggers, and damage to the skin barrier are all factors that make up eczema and because these vary per person, treatment becomes difficult to personalize. AI has helped address this problem, but it often lacks transparency and contextual understanding, which limits how much doctors can rely on it. In a recent study Chourasiya et al. developed an explainable AI framework specifically for eczema that combines different types of data, including electronic health records, skin images, and physician notes, to improve existing AI models and create more personalized treatment plans. The model was trained using a large dataset of over 50,000 dermatology records and analyzed information such as images of the skin, patient medical history, environmental factors, and written clinical notes. It used different models together, including CNNs for images, natural language processing for text, and time-based models for tracking patient history. An important feature of this system is that it uses explainable AI methods like SHAP, which show which factors influenced the model's predictions. The model was tested by comparing its predictions to real patient outcomes and by evaluating it against older ML models, while dermatologists reviewed how useful it was in clinical decision-making. The results showed that the AI performed with an accuracy of about 91.2% and could accurately identify high-risk patients. Doctors also reported a 35% increase in confidence when making treatment decisions and a 27% decrease in misdiagnosis.[25] The introduction of explainable AI frameworks represents an important development in clinical applicability, as interpretability becomes important for algorithmic predictions and physician trust in real-world settings.

Prescription of medication is another challenge clinicians face when treating eczema because not all patients respond the same way to different medications. Researchers Murphy, et al. addressed this problem by developing a ML model to predict whether patients will respond well to Dupilumab, a popular eczema medication, before other treatment begins. The researchers used pretreatment skin biopsy samples from a clinical group of patients with moderate-to-severe eczema who were about to start Dupilumab treatment, along with gene expression data taken directly from their skin tissue. The model analyzed patterns in immune-related genes and inflammation pathways to identify biological differences between patients who later responded well to treatment and those who showed little or no improvement. It was trained using labeled outcomes from past patients; the algorithm learned from real clinical results where doctors already knew whether dupilumab worked or failed. The model was then tested to evaluate how well it could generalize to new patients. The results showed that the model could distinguish responders from non-responders with an AUC of approximately 0.80 to 0.85, which indicates good predictive accuracy, along with balanced sensitivity and specificity for identifying likely treatment failure. [26] This study demonstrated that AI may improve eczema treatment by helping predict nonresponse to Dupilumab early, which could reduce time spent on ineffective treatments

and support more personalized and efficient treatment decisions.

Eczema treatment also often relies on visual assessment and patient-reported symptoms, both of which are subjective, to determine whether a treatment is actually working. This makes it challenging to develop accurate conclusions about the progression of the disease. A 2025 study indexed in PubMed used an AI system to track how eczema patients respond to treatment over time using a combination of subjective and objective indices. The researchers used a group of patients who had been diagnosed with eczema and undergoing treatment, from which they collected standardized facial skin images at different stages of treatment. The model used was a facial image-based deep learning system designed to analyze changes in skin redness, inflammation, and lesion appearance to determine whether a patient was improving or not. The system was also designed to track treatment response over time and provide more consistent monitoring than traditional clinical scoring methods. Dermatologists first evaluated the labeled images for disease severity and treatment response, allowing the AI system to learn visual patterns associated with eczema improvement or worsening It was then validated by comparing its predictions with clinical assessments and standard eczema severity scores used by dermatologists during follow-up visits. The results showed agreement between the AI system and clinician ratings, with correlation values around 0.80 or higher with standard eczema severity scoring systems, showing that the model matched how dermatologists evaluated improvement over time. The system also demonstrated consistent performance across repeated patient visits, suggesting it could reliably track changes in disease severity during treatment rather than just a single snapshot.[27]

A current challenge in eczema management is the connection between diagnosis of the condition, and coming up with personalized treatment suggestions. To address this gap, this 2025 JSIAR study introduced an AI-driven dermatology framework that stands out by integrating both skin disease classification and a recommendation system for targeted treatment. The system works in two main stages: first, a CNN processes input images of the skin to detect and classify the condition, and then once eczema or another condition is identified, the model connects the diagnosis to a structured database of skincare products and treatment options. The model was trained on a dataset of approximately 10,000 labeled dermatoscopic images. After this validation process confirmed that the model could reliably identify skin conditions, the system was then used to generate treatment recommendations, such as using fragrance-free moisturizers, avoiding known irritants like harsh soaps, and choosing products with ingredients suited to the patient's skin sensitivity and severity level. The model achieved an accuracy of about 91.6% and an F1-score of 0.89. Dermatologists also indicated a 35% increase in confidence for treatment recommendations. Combining the two actions into the same model provided more benefits in consistency, helping clinicians come to an accurate conclusion compared to the results if they had chosen to use two different models.[28] Thus, combining AI-based diagnosis with treatment generation may improve eczema management by providing more accessible and individualized care, and could help clinical decision making.

Taken together, AI-based treatment planning systems show potential to improve

personalized eczema care by helping with treatment selection, medication-response prediction, and monitoring of outcomes. However, these technologies are still developing and require larger clinical trials and real-world validation before they can be widely used in dermatology.

Study	AI Method	Dataset	Key Results
Explainable AI Framework (2025)	Multimodal AI + SHAP	50,000+ dermatology records	91.2% accuracy; increased physician confidence by 35%
Dupilumab Response Prediction	Machine learning on gene expression data	Moderate-to-severe eczema patients	Predicted treatment response with AUC of 0.80–0.85
Treatment Response Monitoring (2025)	Deep learning image analysis	Serial patient facial images	Strong agreement with clinician assessments ($r \approx 0.80+$)
AI Diagnosis and Recommendation System (2025)	CNN + treatment recommendation engine	~10,000 skin images	91.6% accuracy; generated personalized treatment recommendations

Section 3 Discussion Section

The studies reviewed throughout this paper collectively suggest that AI has strong potential to improve eczema care by making diagnosis, monitoring, and treatment more accessible, standardized, and personalized. Across nearly all areas of research, the dominant methodology is the use of deep learning systems trained on clinical photographs of eczema lesions. Most of these image-based AI models achieved relatively high performance, often reporting accuracies above 85%, demonstrating that AI systems are capable of identifying visual patterns associated with eczema such as redness, inflammation, scaling, and skin texture changes. Compared to traditional clinical evaluation, which depends heavily on physician experience and subjective interpretation, AI models may help reduce variability between clinicians and improve consistency in eczema-related decision making.

One major strength of current AI systems is their ability to create earlier and more accessible care. Many patients with eczema do not have regular access to dermatologists because of geographic barriers, long wait times, or financial limitations. Several studies addressed this issue through smartphone applications and remote monitoring systems that allow patients to upload images of their skin and receive rapid assessments of disease presence or severity. These approaches could be particularly useful in underserved regions where access to dermatology specialists is limited. Studies involving systems such as EczemaNet, EczemaNet2, and AI-powered mobile applications demonstrated that AI models can perform similarly to dermatologists when evaluating eczema severity, suggesting that these technologies may eventually function as supportive clinical tools for both patients and healthcare providers.

However, most researchers emphasized that these systems are intended to assist clinicians rather than replace them entirely.

Despite promising results, the studies also reveal several major limitations that appear consistently. One of the most significant issues is the lack of diversity in training datasets. Studies largely used datasets dominated by light skin tones, often with only 8% or fewer images representing dark skin.[29] This is a major gap because although individuals of all skin colors, races and ethnicities can be affected by eczema, African Americans, especially children, are more likely to have eczema and it tends to be more severe than those with lighter skin tones.[30] Since erythema and redness may be less visible in deeply pigmented skin, models trained predominantly on fair skin may not generalize well to broader patient populations. Only a small number of studies directly focused on darker skin types, despite the fact that underrepresentation creates substantial risk for reduced diagnostic accuracy and algorithmic bias. Similarly, many datasets lacked sufficient pediatric representation even though eczema is extremely common in children. Some systems, especially for severity and symptom monitoring, specifically focused on pediatric populations, but many studies either concentrated on adults or failed to clearly report age distribution. These limitations reduce confidence regarding how reliably current AI systems would perform across diverse real world populations.

Another important issue is the lack of standardization between studies. Different researchers used different scoring systems, datasets, labeling methods, and outcome measures, making direct comparison between AI models difficult. For severity assessment alone, studies relied on a wide range of clinical scoring systems including EASI, SCORAD, POEM, TIS, and vIGA. Since these scales measure slightly different aspects of eczema severity and may vary between clinicians, it becomes difficult to determine whether one AI model truly performs better than another. Some studies also relied heavily on dermatologist-generated labels as the "ground truth," even though dermatologist scoring itself may vary significantly between clinicians. This highlights a larger issue within AI dermatology research: many systems are evaluated against subjective standards rather than fully objective biological measurements.

Many studies also lacked important information that likely influences eczema severity and treatment response. Environmental factors such as climate, humidity, pollution exposure, seasonal changes, and allergen levels are known to affect eczema symptoms, yet most current AI models do not incorporate these variables into predictions. Behavioral and lifestyle factors such as stress, sleep quality, skincare habits, diet, and treatment adherence are also rarely integrated into existing systems. Several studies acknowledged that future AI models may become more accurate if they combine environmental, behavioral, and clinical data rather than relying primarily on image analysis alone. This may be especially important for predicting flare-ups and creating personalized treatment recommendations, since eczema severity often changes in response to multiple interacting triggers.

Another major limitation involves explainability. Many deep learning systems operate as "black box" models that produce predictions without clearly explaining how decisions were made. This lack of transparency may reduce clinical trust and limit implementation in healthcare

settings where physicians need to understand why a model generated a particular recommendation before incorporating it into treatment decisions.

Long-term implementation also remains an unresolved challenge. Most studies were conducted using static datasets collected over relatively short time periods, meaning it is still unclear how these AI systems would perform over years of real-world clinical use. Eczema presentation, imaging technology, treatment strategies, and patient populations may all change over time, potentially reducing model performance if systems are not continuously updated using new data. Future AI systems will likely require ongoing training through large-scale clinical databases, mobile health platforms, and collaborative hospital networks to maintain reliability and reduce model drift. Continuous data collection may also improve dataset diversity and allow models to adapt to changing patient populations and treatment approaches over time.

Beyond technical limitations, regulatory and ethical concerns must also be considered. Many AI systems require large collections of patient photographs, medical records, and genomic data, raising concerns about privacy and data security. Patients may be reluctant for their information to be public use and would choose to keep it private, making it hard to create training data. Questions also remain regarding clinical responsibility when AI-generated recommendations conflict with physician judgment. Regulatory agencies are still developing frameworks for evaluating and approving AI-based medical technologies, particularly systems that continuously learn from new data. Addressing these concerns will be essential before AI can be fully integrated into routine eczema care.

Future research should focus on conducting large, prospective clinical studies that evaluate AI systems under real-world conditions rather than relying primarily on retrospective datasets. Greater collaboration among hospitals, research institutions, and technology developers could facilitate the creation of larger, more diverse datasets that better represent different skin tones, age groups, geographic regions, and disease severities. Standardizing evaluation metrics and clinical scoring systems would also allow researchers to compare AI models more consistently across studies. In addition, future models should incorporate multimodal data, including clinical photographs, electronic health records, genomic information, environmental exposures, and patient-reported outcomes, to better reflect the complex and multifactorial nature of eczema. Finally, continued research should prioritize explainable AI, privacy protection, and regulatory validation to ensure these technologies are transparent, trustworthy, and suitable for routine clinical use.

Overall, the studies suggest that AI assisted eczema care is progressing rapidly and may significantly improve multiple aspects of dermatology in the future. Future research should focus on developing multimodal AI systems that combine imaging, molecular data, patient history, behavioral factors, and environmental information in order to create more comprehensive and personalized eczema management systems. Although many of these technologies remain in relatively early stages of development, the findings reviewed throughout this paper strongly suggest that AI may eventually become an important supportive tool within dermatology and significantly improve the management of eczema across diverse patient populations.

Conclusion

AI is quickly becoming an important tool in dermatology, particularly in the diagnosis, monitoring, and treatment of eczema. The studies reviewed demonstrate that AI systems can accurately identify eczema from clinical images, assess disease severity, monitor symptoms over time, predict flare-ups, and assist in developing more personalized treatment plans. Many of these systems achieved performance comparable to trained dermatologists, suggesting that AI can improve both the efficiency and accuracy of clinical care.

At the same time, several challenges remain before AI can be widely implemented in everyday eczema care. Many current models are trained on limited datasets that do not adequately represent diverse skin tones, age groups, or environmental conditions. Differences in study methods and evaluation criteria also make it difficult to compare results across models. Concerns about explainability, long-term reliability, and integration into healthcare systems must also be addressed.

Despite these limitations, the overall findings suggest that AI can play an important role in the future of eczema management. As datasets become larger and more diverse, AI may provide more accurate and personalized care. While AI is unlikely to replace dermatologists, it can serve as a valuable tool that supports healthcare professionals by enabling earlier diagnoses, improved monitoring, and more effective treatment.

Works Cited

Cleveland Clinic. Eczema: What It Is, Symptoms, Causes, Types & Treatment. Cleveland Clinic, https://my.clevelandclinic.org/health/diseases/9998-eczema. Accessed 22 July 2026.

National Eczema Association. Eczema Facts. National Eczema Association, https://nationaleczema.org/eczema-facts/. Accessed 22 July 2026.

Michigan Technological University. What Is Artificial Intelligence (AI)? Michigan Technological University, https://www.mtu.edu/data-science/undergraduate/ai/what-is/. Accessed 22 July 2026.

Naik, Pravin P. "Treatment-Resistant Atopic Dermatitis: Novel Therapeutics, Digital Tools, and Precision Medicine." Asia Pacific Allergy, vol. 12, no. 2, 2022, article e20. https://doi.org/10.5415/apallergy.2022.12.e20.

National Eczema Association. Eczema Treatments. National Eczema Association, https://nationaleczema.org/treatments/. Accessed 22 July 2026.

Cao, F., et al. "Advancements in Artificial Intelligence for Atopic Dermatitis: Diagnosis, Treatment, and Patient Management." Annals of Medicine, vol. 57, no. 1, 2025, article 2484665. https://doi.org/10.1080/07853890.2025.2484665.

Clark, Ben. "What Is Image Classification?" IBM, 30 Sept. 2025, https://www.ibm.com/think/topics/image-classification.

Image Classification Using Convolutional Neural Networks (CNNs). ResearchGate, 2025, https://www.researchgate.net/publication/396609546_Image_Classification_using_Convolutional_Neural_Networks_CNNs.

Rasheed, A., et al. "Automatic Eczema Classification in Clinical Images Based on Hybrid Deep Neural Networks." Computers in Biology and Medicine, vol. 147, 2022, article 105807. https://doi.org/10.1016/j.compbiomed.2022.105807.

Hanifin, Jon M., et al. "A Population-Based Survey of Eczema Prevalence in the United States." Dermatitis: Contact, Atopic, Occupational, Drug, vol. 18, no. 2, 2007, pp. 82–91. https://doi.org/10.2310/6620.2007.06034

Kalbande, D., et al. "An Artificial Intelligence Approach for the Recognition of Early Stages of Eczema." International Journal of Medical Engineering and Informatics, vol. 12, no. 1, 2020, p. 52. https://doi.org/10.1504/IJMEI.2020.105656.

Ding, N., et al. "Differential Diagnosis of Eczema and Psoriasis Using Routine Clinical Data and Machine Learning: Development of a Web-Based Tool in a Multicenter Outpatient Cohort." Frontiers in Medicine, vol. 12, 2025. https://doi.org/10.3389/fmed.2025.1667794.

Liu, H. M., et al. "Applying Artificial Intelligence to Identify Common Targets for Treatment of Asthma, Eczema, and Food Allergy." International Archives of Allergy and Immunology, vol. 185, no. 2, 2023, pp. 99–110. https://doi.org/10.1159/000534827.

Ajani, A., et al. "Artificial Intelligence-Assisted Diagnosis of Atopic Eczema in Darker Skin Types: A CNN-Based Study." Dermatology Practical & Conceptual, vol. 16, no. 1, 2026, article 5897. https://doi.org/10.5826/dpc.1601a5897.

Kaltwasser, Jill. "AI Model Can Help Classify Atopic Dermatitis Lesion Severity." The American Journal of Managed Care, 29 May 2025, https://www.ajmc.com/view/ai-can-help-classify-atopic-dermatitis-lesion-severity.

Hurault, G., et al. "Personalized Prediction of Daily Eczema Severity Scores Using a Mechanistic Machine Learning Model." Clinical & Experimental Allergy, vol. 50, no. 11, 2020, pp. 1258–66. https://doi.org/10.1111/cea.13717.

Zvulunov, A., et al. "A Mobile Health App for Facilitating Disease Management in Children with Atopic Dermatitis: Feasibility and Impact Study." JMIR Dermatology, vol. 6, no. 1, 2023, article e49278. https://doi.org/10.2196/49278.

Fong, W. C. G., et al. "Can Artificial Intelligence-Powered EczemaNet Facilitate Whole-Person Remote Eczema Assessments Using Digital Images?" British Journal of Dermatology, vol. 193, suppl. 1, 2025, article ljaf085.036. https://doi.org/10.1093/bjd/ljaf085.036.

Attar, R., et al. "Reliable Detection of Eczema Areas for Fully Automated Assessment of Eczema Severity from Digital Camera Images." JID Innovations, vol. 3, no. 5, 2023, article 100213. https://doi.org/10.1016/j.xjidi.2023.100213.

Bibi, I., et al. "Automated Machine Learning Analysis of Patients with Chronic Skin Disease Using a Medical Smartphone App: Retrospective Study." Journal of Medical Internet Research, vol. 25, no. 1, 2023, article e50886. https://doi.org/10.2196/50886.

Ouroutzoglou, M., et al. Quantifying Itch and Its Impact on Sleep Using Machine Learning and Radio Signals. arXiv, 2025, https://doi.org/10.48550/arXiv.2501.04896.

Xing, Y., et al. "Quantifying Nocturnal Scratch in Atopic Dermatitis: A Machine Learning Approach Using Digital Wrist Actigraphy." Sensors, vol. 24, no. 11, 2024, article 3364. https://doi.org/10.3390/s24113364.

Yang, A. F., et al. "Artificial Intelligence-Enabled Wearable Devices and Nocturnal Scratching in Mild Atopic Dermatitis." JAMA Dermatology, vol. 161, no. 4, 2025, pp. 406–10. https://doi.org/10.1001/jamadermatol.2024.5697.

Nielsen, M.-L., et al. "Predictors of Flares and Disease Severity in Patients with Atopic Dermatitis Using Machine Learning." JAMA Dermatology, vol. 161, no. 9, 2025, pp. 950–56. https://doi.org/10.1001/jamadermatol.2025.2073.

Chourasiya, L., S. Khatri, and K. K. Sethi. "AI-Driven Personalized Care in Atopic Dermatitis: Integrating Explainability for Clinical Adoption." Proceedings of the IEEE International Conference on ICT in Business Industry and Government (ICTBIG 2025), IEEE, 2025. https://doi.org/10.1109/ICTBIG68706.2025.11323645.

Murphy, M. J., et al. "Machine Learning Analysis of Pretreatment Skin Biopsies Predicts Nonresponse to Dupilumab in Patients with Eczematous Dermatitis." British Journal of Dermatology, vol. 190, no. 1, 2023, pp. 132–34. https://doi.org/10.1093/bjd/ljad389.

Parraga, S. P., et al. "Using Facial Skin Artificial Intelligence (AI) to Capture Atopic Dermatitis Patients' Response to Therapy: A Single-Center Prospective Study." International Journal of Dermatology, vol. 64, no. 8, 2025, pp. 1509–11. https://doi.org/10.1111/ijd.17688.

Haque, M. M., M. S. Rahman, and M. R. Islam. AI-Driven Diagnosis and Management Framework for Atopic Dermatitis Using Machine Learning Techniques. Journal of Scientific Innovation and Applied Research, 2025, https://jsiar.com/2025-May/JSIAR-M-25-05800.pdf.

Shabu, A. M., M. S. Ansari, and A. Aslam. A Generative AI Approach for Reducing Skin Tone Bias in Skin Cancer Classification. arXiv, 2026, https://arxiv.org/abs/2602.14356.

National Eczema Association. Eczema Facts. National Eczema Association, https://nationaleczema.org/eczema-facts/.

The Nature and Nurture of Schizophrenia: A Review of Gene–Environment Interactions
By Gurkriti Kochhar

Abstract

Schizophrenia, a complex psychiatric disorder that devastates 1% of the human population, has long been regarded as a disorder of high heritability due to extensive research into its genetic pathways, especially with recent instances of intensive genotyping through genome-wide association studies (GWAS). GWAS have uncovered a number of genetic contributors within schizophrenia, categorized both as common variants (single nucleotide polymorphisms) that hold limited individual risk, and rare variants (copy number variations) that pose high risks to development. However, simultaneously to genetic focused heritability, environmental factors have also been shown to play a role within the disorder, which can be categorized as psychosocial (e.g, trauma), biological (e.g, substance abuse), and physical factors (e.g, prenatal complications). With the prevalence of both genetic and environmental factors being proven, their interaction is a topic of great importance within this field. In this review, we examine the roles of individual genetic and environmental factors in the pathology of schizophrenia, in addition to specific case studies of the interactions between the two in the development of psychosis. Finally, we will discuss the convergence of these factors to understand specific molecular pathways, highlighting that such studies are vital to tackle the disorder's complexity.

1. Introduction

Schizophrenia, while widely studied, is still far from understood. It is a hugely devastating, complex chronic mental disorder that impacts around 1% of our worldwide population, with a usual age of onset between 16 and 30 years (Réthelyi et al.; *Schizophrenia*). This disorder is characterized by a consistent and recurring state of psychosis, which results in severe changes to the manner in which its patients think, behave, and perceive the reality around them (*Schizophrenia*). Symptoms can include hallucinations, impaired speech, social withdrawal, lack of emotions, consistent suspicion, disorganized speech, apathy and lack of executive functioning (Patel et al.). Beyond these clinical symptoms, schizophrenia carries significant social implications. Patients with schizophrenia often may not believe that they are experiencing any sort of psychosis, making those who deal with the illness often targets of societal taboo (*Schizophrenia*). Not only does this dull their prospects for improvement, but their lack of touch with reality can make daily life difficult for the patients and even their families (*Schizophrenia*).

Apart from the disorder's troubling circumstances, its etiology is an even harder hill to climb. Schizophrenia is inherently heterogeneous, making it unlikely to find one root cause for the disorder across every patient (Patel et al.). From a genetic standpoint, schizophrenia is seen as highly heritable, with a commonly accepted prevalence of genetic risks sitting at around 80% (Sullivan et al.). However, as more investigation is carried out, it has been noticed that many environmental conditions, such as trauma, prenatal complications, or advanced parental age,

have showcased their own positive correlation with the pathophysiology of schizophrenia (Réthelyi et al.). As such, a vital discussion about the interplay between genetic and environmental factors has discouraged the view that these factors are sole determinants. With the help of family studies and animal models, there are increasing efforts in trying to determine the paths of pathophysiology, in order to ensure that more patients with schizophrenia can be appropriately treated.

In this review, we discuss a variety of factors related to schizophrenia. We will focus on specific common and rare genetic mutations that have been shown to impact the pathophysiology of schizophrenia. We then discuss the physical, biological, and psychological aspects of how the environment implicates risk to schizophrenia. Lastly, we look into the interactions between the nature (genetic) and nurture (environmental) of schizophrenia, and how their relationship can aid in future treatment of this disorder.

2. Overview of Schizophrenia

In 1908, Professor Eugen Bleuler devised the term "Schizophrenia" to describe a mental disorder that created a "split" in one's mind. This term originates from the disorder's root in distorted perception and a lack of unity in psychological functions, rather than a split personality (Ashok et al.). Schizophrenia is characterized by being an early-onset chronic disease that creates a cluster of positive symptoms as well as negative symptoms. Positive symptoms include hallucinations, impaired speech patterns, delusions, and irregular behaviors (Patel et al.), while negative symptoms refer to withdrawal from daily activities, lack of pleasure, and a lack of emotion in regular speech and expressions (often associated with flat affect) (D'Arrigo and Ellis). These symptoms all come together to make daily life difficult for those who are diagnosed. Life expectancy dips anywhere from 10-20 years, and often, only 10-25 % of diagnosed individuals are employed steadily (Owen et al.).

Schizophrenia is often prevalent in people who have a history of mental illness, caused either by their genetic background, environmental background, or a combination. Although the exact cause of schizophrenia is still unknown, most theories suggest that the disorder is associated with abnormalities in neurotransmission. There is often either a lack or an excess of neurotransmitters such as dopamine, serotonin, gamma-aminobutyric acid, and others (Yang and Tsai). Issues with these neurotransmitters frequently result from environmental or genetic factors. Risks in one's environment, such as childhood trauma, social isolation, or even substance abuse, can trigger acute stress, leading to imbalances. Additionally, genes pose a threat to heritability, and certain mutations relating to brain development can increase the risk of schizophrenia when combined with a traumatic event (Réthelyi et al.).

Due to the lack of understanding of the precise causes of schizophrenia, a variety of treatments currently exist, in hopes of targeting any possible malfunctions. Psychotherapy aims to complement medication by providing individual counseling, behavioral therapy, and compliance therapy, as well as treatment to help patients re-acclimate to group settings. This type

of therapy even works toward keeping patients adherent to pharmaceutical approaches, ensuring that they don't deny their illness and fall into relapse (Patel et al.).

3. Genetic Factors

The etiology of schizophrenia cannot be pinpointed to a specific pattern of genetic variation (Réthelyi et al.). Rather, our current knowledge of the genetic risk factors of schizophrenia points to its polygenic nature, with its risk alleles being through multiple DNA variants at differing loci (Rees et al.). In this chapter, we will discuss the heritability of schizophrenia as supported by twin studies, as well as the most prevalent single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) relating to the etiology and pathophysiology of schizophrenia.

3.1 Twin studies

Results from schizophrenia twin studies have consistently served as a validation of genetic factors playing into its etiology (Kendler). Evidence from a synthesis of 5 twin studies from 1995-2002 conducted in Europe and Japan showcased concordance rates of 41-65% in monozygotic twin pairs (100% of common genes) and between 0-28% in dizygotic pairs (50% of common genes) (Cardno and Gottesman). A study based on the nationwide Danish twin register held concordance rates for MZ and DZ twins to 33% and 7% respectively, with another individual study showing rates of 50%, and 17% respectively (Modinos et al.; Hilker et al.). Given the noticeably higher margin in MZ twins as opposed to DZ twins, this data, along with that of another meta-analysis, agrees that the heritability of schizophrenia falls ~80% (Sullivan et al.; Hilker et al.; Cardno and Gottesman). Arguments against the 80% statistic argue that Sulivin's original meta-analysis was conducted with a 'relaxed' criterion of inclusion, but the metric has been backed with alternate studies (Joseph; Hilker et al.; Cardno and Gottesman). Despite this conclusion, it is still important to acknowledge that this disorder is still far from being fully genetically determined, and is rather just influenced by these factors.

3.2 SNP

One category of the genetic factors that influence schizophrenia is SNPs. SNPs are also often called the "common" sources of genetic variation due to how they denote a mutation in a single base pair within the DNA (Lakna). Due to this, they have a relatively small impact (common odds ratio (OR) of <1.2) on schizophrenic prevalence, with their true impacts being additive (Rees et al.; Réthelyi et al.). Recent genome-wide association studies (GWAS) have been used to determine the statistical impact of many SNPs relating to schizophrenia due to their access to a large sample size (Schizophrenia Working Group of the Psychiatric Genomics Consortium).

3.2.1 GRM3

One of the genes implicated in having an association with SNPs in schizophrenia by the GWAS is glutamate receptor 3 (GRM3). The presence of dimerized mGluR3 is thought to be reduced in patients with schizophrenia, with polymorphisms on the gene resulting in schizophrenia associated changes such as cognitive impairment and changes in glutamate signaling (resulting in alternative brain activation patterns and altered development) (P. Harrison et al.). In a meta-analysis on 14 SNPs in the GRM3 gene across 11318 schizophrenia cases, 2 SNPs (rs2237562, rs917071) had a strong linkage disequilibrium (connection in same-time prevalence) as rs12704290, which was labeled the top significant G3M3 SNP by the Schizophrenia working group of the Psychiatric Genomics Consortium (PGC), a group that worked to put together 108 schizophrenia associated loci (Schizophrenia Working Group of the Psychiatric Genomics Consortium; Saini et al.). However, it is worth noting that these SNPs were only prevalent depending on the population in study, denoting that the severity of GRM3 risk is ancestry dependent (Saini et al.).

3.2.2 MIR137

Another target gene identified by the PGC is microRNA-127 (MIR137), specifically with a SNP on rs1625579 within the European population, but with limited to no impact within Asian populations (Liu et al.; Pu and Xiao). MIR137 has been shown to be involved in neural development, particularly in developmental processes and responses to stimuli (Bodai et al.). Patients having the rs1625579 SNP have cognitive impairments with an increase in negative symptoms such as reduced emotion, lack of motivation, and social withdrawal (Réthelyi et al.).

3.2.3 COMT

Another gene of interest with SNPs is catechol-O-methyltransferase (COMT), which is involved in the breakdown of catecholamines such as dopamine (Modinos et al.). A key and highly studied polymorphism with COMT is COMTval158met or polymorphism rs4680 (J. Ma et al.). Studies about COMT typically have varied outcomes, with some negative results of no association (Modinos et al.), however, through a family-based association analysis, high expression of the val allele in COMT has been studied to increase the incidence of schizophrenia due to its role in increasing COMT expression, which in turn results in lower dopamine levels, weakening prefrontal cortex activity (Egan et al.).

3.2.4 DISC1

Lastly, the most researched gene in schizophrenia, Disrupted in Schizophrenia 1 (DISC1), has implications of SNPs in the development of the disorder (Doctoral School, Department of Neurosciences, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania et al.; J.-H. Ma et al.; Cardno and Gottesman). Within its regular functioning, the DISC1 gene encodes for an intracellular scaffold protein that organizes neuronal migration, intracellular transport, and maintenance of synapses. Disruptions within this gene can alter brain structure and impair prefrontal cortex function, resulting in some DISC1 SNPs increasing schizophrenia risk

(Doctoral School, Department of Neurosciences, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania et al.). In an Iranian population study, 3 particular SNPs, rs6675281, rs2255340, and rs2738864, showcased OR: 2,629, OR: 1.897, OR: 3.605 respectively in their prevalence in schizophrenia risk within the Iranian population (Shokouhifar et al.).

3.3 CNV

The second category of genetic factors that result in schizophrenia is CNVs. CNVs are a "rare" type of genetic variation in the genome because of how they involve whole sections of the genome being deleted or duplicated, rather than single-point mutations (Réthelyi et al.; Lakna). Their rarity also denotes their high risk, with CNVs having a large impact on the etiology of schizophrenia (OR between 2 and 60) (Réthelyi et al.; Rees et al.). In this review, we will discuss certain microdeletions (with decreased or absent expression due to deletion of the copies).

3.3.1 22q11.2 Deletion

One of the most common CNVs studied with schizophrenia is the microdeletion of 22q11.2, with a prevalence of around 25-30% of patients with the deletion developing schizophrenia, and 0.9% of total patients with schizophrenia having the mutation (Qin et al.; Van et al.). In an animal model of the deletion, mice that were heterozygous for the deletion showcased learning defects, memory impairment, and significantly reduced prepulse inhibition, all being typical symptoms of patients with schizophrenia due to a lowered sensory function (Qin et al.; Paylor). Other distinguishing features of the deletion include alternative physical features (dysmorphic), lower IQ, and mental retardation (Bassett and Chow). Commonly, schizophrenia is associated with a 3 million base (Mb) pair deletion of the 22q11.2, of which a large majority are de novo mutations that occur during sperm or egg formation (new cases show 5-10% of inherited deletions) (Bassett and Chow).

3.3.2 1q21.1 Deletion

Another primary CNV is the microdeletion of 1q21.1 in schizophrenia, showcasing a prevalence of an 8-fold risk of development, with 0.2-0.65 of patients with schizophrenia showcasing the deletion (Réthelyi et al.; Guo et al.; Szecówka et al.). 1q21.1 microdeletions are associated with cognitive impairment driven by structural changes (often reductions) to the brain (Guo et al.). Such reductions include agenesis of the corpus callosum (resulting in difficulty processing and integrating information), reduced hippocampal volume (linked with emotional dysregulation and cognitive impairment), and reduced volume in subcortical structures overall when comparing carriers of this deletion (Guo et al.). 1q21.1 deletions typically show a length from 860 kb to 2.8 Mb, and have a lower presence of de-novo mutations as compared to 22q11.2 deletions at less than 40% (Bassett et al.).

3.3.3 15q13.2-3

Lastly, the microdeletion on 15q13.2-3 showcases a prevalence in schizophrenia similar to that of a 1q21.1 microdeletion, with around 0.2-0.3% of patients with schizophrenia holding the mutation (Bassett et al.). In a report characterizing a 15q13.3 knockout animal model showcased seizure susceptibility, auditory sensory inhibition, and reduced prepulse inhibition in a gene-dosage-dependent manner (Forsingdal et al.). Additional changes included deficits in spatial memory (driven by cognitive and visual impairments), as well as gene-dosage-dependent reduction in nest-building activity, which is characteristic of negative symptoms such as reduced motivation, a state of apathy, and social withdrawal (Forsingdal et al.). 15q13.2-3 deletions are often between 500 kb and 3.8 Mb, typically around 1.5 Mb, where (similarly to 1q21.1 deletion syndrome) around <40% of the mutations appear de novo (Bassett et al.).

4. Environmental Factors

Going alongside the genetic risk factors previously discussed relating to the etiology of schizophrenia, research into the disorder has also been targeting the environmental aspect of the development (Réthelyi et al.), In this section, we will go through these factors by branching them under 3 main subcategories, which include: psychosocial, biological, and physical environmental factors. Psychosocial factors include trauma, migration, and social isolation. Biological factors include cannabis and other substance use. Lastly, physical factors include prenatal complications and advanced parental age.

4.1 Psychosocial factors

There have been many investigations into how psychosocial factors can play a part in being a risk factor for schizophrenia. A meta-analysis focusing exclusively upon childhood trauma revealed a significant spike in psychosis (OR = 2.78) through external trauma as defined by sexual abuse, physical abuse, emotional/psychological abuse, neglect, bullying, and parental death (Varese et al.). This evidence also implies a high prevalence of trauma in the current, as supported by a population-based study where patients who underwent sexual and physical abuse displayed a higher reporting of hallucinatory behaviours (OR = 1.7) (Morgan and Fisher). In addition to traumatic childhood events, migration can also have a severe impact on symptom appearance (Morgan and Fisher). Increased risk with both second and first generation immigrants remains relatively constant, with no statistical risk difference (Bourque et al.). Such risks have been studied with Danish immigrants (OR=1.93 for one foreign-born parent and OR=2.96 for both foreign parents) (Cantor-Graae and Pedersen), Germany (G. Harrison, Glazebrook, et al.), and other high-income countries such as the UK (Stilo and Murray). Lastly, social inequality and isolation that come with class differences can also play a role as a risk factor, however, with significantly less proof than that of childhood ailments (Beards et al.). However, in childhood, economic instability of guardians is also associated with higher risk, with those who were born in deprived areas at birth showcasing an increase in adult-onset schizophrenia (OR=8.1) (G. Harrison, Gunnell, et al.). Patients with schizophrenia also showcase a higher probability of living alone and receiving a lower income (Stilo and Murray).

4.2 Biological factors

Cannabis is a widely consumed drug, with around 120 to 224 million users worldwide. This drug can be cited to have an effect of well-being or elation, but in higher dosage, can result in mental illness or psychosis, including schizophrenia (McLoughlin et al.). While 16% of the general population has issues with drug and alcohol abuse, that percentage significantly increases to 47% when concerning patients with schizophrenia (Khokhar et al.). In addition, substance users displayed more positive symptoms (hallucinations, delusions, irregular speech, etc.), which require more frequent hospital visits and closer care (Swofford et al.). In a study of Swedish conscripts, there was a higher observed rate of developed schizophrenia among those who had been using cannabis only (OR=1.9). Additionally, there was an increased amount of correlation between the development of schizophrenia and higher dosage of cannabis used (Zammit et al.). Finally, patients with a history of cannabis use have more frequent and longer hospital stays, indicating that the utilization of cannabis can attenuate symptoms of psychosis (Pearson and Berry).

4.3 Physical factors

Lastly, physical factors such as prenatal complications and advanced parental age are other environmental factors that influence one's susceptibility to schizophrenia. In a study examining the correlation between parental age and the risk of schizophrenia in adult offspring, there is a consistent increase in the percentage of cases in which the offspring displays schizophrenia spectrum disorders as parental age increases, going from 1.13 with a parental age of 15-24, to 1.83 occurrences with a parental age of 45-68 (Brown et al.). Another study conducted in Sweden emphasized a correlation specifically between the father's age and the odds of the offspring developing schizophrenia, in which there was a high relation between the development of the disorder and an advanced paternal age of 45 or higher (when compared to aged 20-24, OR=2.8) (Dalman and Allebeck). Unlike paternal age, maternal age has a much more minimal impact, with there being weak to absolutely no association observed in studies, especially with paternal ages controlled for (Dalman and Allebeck; Sipos et al.). However, this conflicts with the findings of Lopez-Castroman, which saw a significant linear increase in psychiatric disorder in offspring as maternal age advanced (Lopez-Castroman et al.). Higher parental age is most notably attributed to de novo mutations that arise in paternal germ cells as the father ages (Malaspina et al.). Other than parental age, prenatal complications have often been physical risk factors associated with the appearance of schizophrenia. During a stressful pregnancy, several risk factors could occur, such as maternal denutrition, exposure to influenza, excessive bleeding, preeclampsia, emergency C-section, reduced head circumference, low birth weight, and repeated psychological stress (Boog). The risk for schizophrenia is doubled in the instance of prenatal complications (Boog).

5. Gene x Environment Interactions

Now that the individual risks that genetics and the environment play in the prevalence of schizophrenia have been discussed, it is vital to understand the interplay of these two factors. As proven by the twin studies mentioned when discussing genetic implications on schizophrenia, genetics plays a large part, but not the sole part, in the etiology (Modinos et al.). Most notably, studies of discordant MZ twins show that heritability is not the sole factor in development (Cardno and Gottesman). In addition, the universal prevalence of many aforementioned environmental factors, such as cannabis use or trauma, is high, but only a few people develop schizophrenia, exposing the truth that there needs to be collaboration between nature and nurture (Réthelyi et al.). In the following portions, we will review some of the pathways of development of schizophrenia relating to genes and the environment.

5.1 Interactions of Cannabis use and COMT

The interaction between cannabis use and COMT is one of the most studied pathways (Wahbeh and Avramopoulos). The role that the COMT gene plays in dopamine breakdown makes it a moderator for psychosis development with the use of cannabis (Wahbeh and Avramopoulos; Grechuk et al.; Ruud Van Winkel et al.). In a study conducted on a sample of 135 patients with schizophrenia who were genotyped for the COMT Val158Met SNP, cannabis use was monitored on the basis of history of abuse, measuring verbal fluency and speed of processing. In the results, there was a noticeable decrease in capability in the patients with the Val/Val genotype, signifying that the impact of cannabis has a more detrimental effect on those with an increase in COMT activity (Bosia et al.). In addition to this study, a paper on the correlation of cannabis use and COMT presence in Turkish men reported that patients who used cannabis and had the Val/Val genotype had significantly higher schizophrenia prevalence (88,9%) as compared to those with the same genotype but no cannabis use (68.4%) (Ermis et al.). The tetrahydrocannabinol (THC) present within cannabis results in an increase of dopamine, which, in combination with the decrease of dopamine within val/val allele COMT SNPs, causes a disadvantage conflict leading to increased stress responses and high vulnerability (Bosia et al.).

5.2 Interactions of Trauma and COMT

As discussed in the previous section, trauma results in a significant increase in one's susceptibility to schizophrenia, and in relation to COMT, a gene that encodes an enzyme that regulates dopamine, psychosis related events can be amplified. However, it is worth noting that the results for such studies are often conflicting. For some individuals who are homozygous for the Met allele at COMT Val158Met, they showcase a higher tendency to stress and a higher propensity to anxiety due to their higher dopamine activity, as well as an increased psychoticism as compared to the control (Wahbeh and Avramopoulos; R. Van Winkel et al.; Mathew et al.). In other studies, however, such as one that noted a group of men entering military training, those with the Val allele showcased a "paranoid ideation" along with psychosis, indicating that stressful moments result in an increased risk of psychosis in both carriers of a homozygous Met and Val allele (R. Van Winkel et al.). This psychosis showcases its link to schizophrenia by being

one of the main positive symptoms in the disorder, associated with a dissociation from reality (Wahbeh and Avramopoulos).

5.3 Infection

Prenatal maternal infections of *Toxoplasma gondii (Toxo)* have been shown to increase the risk of stunted neurodevelopment of the fetus, which has led to it being one of the most researched infections in relation to schizophrenia (Shorter and Miller; Ayhan et al.). Prenatal *Toxo* infections have showcased an 80% increase in one's odds of schizophrenia due to their involvement in increasing dopamine production and the production of inflammatory factors that cause increased aggression and impulsivity (Robinson and Bergen). In addition, the infection showcases schizophrenia-like effects of decreased social cognition, learning ability, and increased negative symptoms such as withdrawal (Guimarães et al.). The gene and environment interaction in the scope of *Toxo* implies that an increased susceptibility to the infection is associated with the same genes that result in the neurodevelopmental issues of schizophrenia (Wang et al.).

5.4 Epigenetics

DNA methylation has been present in multiple neuropsychiatric disorders just such as schizophrenia due to its repression of genes vital to the neural pathway (Shorter and Miller). A study on monozygotic twins (which denotes a common DNA sequence) represents epigenetics as a source of schizophrenia etiology in one twin and not the other. Through a genome-wide analysis, 22 pairs of twins were shown to have their ST6GALNAC1 promoter regions differently methylated when they were discordant for schizophrenia (Dempster et al.). In addition, DNA methylation has been implicated in reduced expression of genes such as Brain-derived Neurotropic Factor (BDNF), COMT, reelin (RELN), Glutamate decarboxylase-67 (GAD1), and others involved in neural development and function pathways (Shorter and Miller; Magwai et al.). Certain pre or post-natal factors can influence the methylation of the genes listed above (Shorter and Miller). Due to a lack of post-natal maternal care, GAD1 promoters can be methylated, which results in modifications that alter adult development, with symptoms such as a decreased ability to regulate anxiety (Shorter and Miller; Bagot et al.). In addition to DNA methylation, histone modifications are also vital to disease pathogenesis . Histone 3 is commonly dysregulated in patients with schizophrenia, due to its reduction of the expression of genes that target neuronal function and metabolism (Shorter and Miller). Increased acetylation of H3 is associated with enhanced function in spatial memory and synaptic plasticity, both of which are commonly impaired in schizophrenia (Shorter and Miller; Chen et al.). These epigenetic factors are important to note, as their reversible nature (compared with mutations in one's actual genome) makes them key targets for regulating genes that can increase the risk of schizophrenia (Thomas).

6. Current and Future Therapies

The treatment of schizophrenia has come a long way from its state at the disorder's realization a century ago (Maric et al.). For individuals impacted by the disorder in the past, many unfocused and heavily inhumane treatments, including deep insulin comas, metrazol shock therapies, electroconvulsive therapy (ECT), and lobectomies, were applied (Maric et al.; LaMarco). Insulin comas (utilized between the 1920s and 1950s) aimed at "resetting" the brain with high dosages of insulin; metrazol therapy (1930s to 1940s) resulted in violent convulsions, while ECT was used to elicit seizures (and is still utilized today with heavier control) (LaMarco). In their previous state, such treatments endangered the patients and added an increased level of stigma around the disorder due to their ethical issues.

However, even as of now, the lack of understanding of the specifics of the etiology and pathology of schizophrenia has resulted in many gaps within the therapeutic solutions for patients and a large amount of the population that the disorder impacts. These gaps can be characterized as 3 main current issues (Stępnicki et al.). Firstly, current therapies are only effective for half of the population of those with the disorder (Stępnicki et al.). Secondly, many current therapies overlook negative symptoms and treat only around half of the patients (Réthelyi et al.; Stępnicki et al.). Lastly, they result in neurological and metabolic side effects (Stępnicki et al.). For the classical treatment approaches to schizophrenia, antipsychotic drugs that antagonize dopamine receptors are utilized (Stępnicki et al.; Yang and Tsai; LaMarco). The first generation of these antipsychotic drugs (typical) worked to block D_2 dopamine receptors without real selectivity as to the part of the brain, which resulted in a range of side effects specifically in motor skills (LaMarco). However, because these first-generation drugs were not selective in their receptor targeting, they frequently resulted in severe muscle breakdown, unwanted movement, tremors, and stiffness due to blockage of D_2 dopamine receptors in the nigrostriatal pathway (Stępnicki et al.; "Chlorpromazine Tablets"). However, as the second generation (atypical) of antipsychotics were introduced, schizophrenia treatments started leaning a lot more towards regulation of other pathways that allowed for a stronger reduction in negative symptoms, and less motor issues within the patients (LaMarco; Stępnicki et al.). The evolution of these drug therapies showcases an increase in knowledge about alternative pathways that lead up to the pathology of schizophrenia.

The prevalence of gene and environmental interactions is quite significant to the future of schizophrenia therapies. An understanding of the genetic pathways, in collaboration with environmental effects, can be used as a preventive mechanism for the disorder among those aware of a hereditary risk. As such, prevention can be achieved with early environmental modifications (Réthelyi et al.). In addition, a main advantage of understanding gene and environmental interactions is the ability to study scenarios in greater detail through animal models (Réthelyi et al.; Modinos et al.). Such models can help determine the efficiency of treatments for patients who have had different collections of factors leading up to their episodes of psychosis, helping researchers understand resistance to treatments and also personalized treatments that depend on family counseling (Réthelyi et al.; Modinos et al.).

While the progress in understanding the gene and environmental interactions in relation to schizophrenia is promising, there are still a lot of limitations in treatments due to the nature of the disorder. As proved by the interactions discussed in the previous section, as well as the high volume of different factors prevalent to schizophrenia's etiology, the disorder has a high complexity that vastly differentiates the pathways that are prevalent from patient to patient. As such, the complexity of the disorder would also benefit from collections of genetic biobanks and additional collaborative efforts (such as the PGC) to optimize the quality of the connections drawn between genetic and environmental factors (Modinos et al.).

Conclusion

As shown by the literature in this review, gene and environment interactions have been a highly prevalent course of study within schizophrenia. Within this review, we identified the root of the complexity relating to schizophrenia as attributed to the influence of both genetic and environmental factors, which denotes that studying the interactions between the two can contribute to our understanding of schizophrenia's development. We emphasized individual genetic factors via common SNPs and rare CNVs, individual environmental factors of psychosocial, biological, and physical nature, as well as case studies and GxE interactions through epigenetics. Lastly, we addressed what these interactions could mean for the future of schizophrenia treatment and how they would differ from the past and current treatment options.

Due to its complexity, current research still lacks a causal model for this disorder, which has thus been unable to serve as a basis for new treatment plans (Réthelyi et al.). However, schizophrenia research has been greatly improved with the study into the interactions between genetics and the environment (Réthelyi et al.; Modinos et al.). Not only is prevention much easier to study for those with high hereditary risk, with the utilization of animal models and biobanks, a higher quality of information can also help in the understanding of patient-specific resistance to current therapies, as well as help with the understanding of pathways alternative to the dopamine hypothesis (Modinos et al.; Yang and Tsai; Hilker et al.). Moving forward, it is vital to develop tools that can compute GWAS findings alongside environmental data to accurately simulate the disorder for study (Réthelyi et al.).

Works Cited

Ashok, Ahbishekh Hulegar, et al. "Paul Eugen Bleuler and the Origin of the Term Schizophrenia (SCHIZOPRENIEGRUPPE)." Indian Journal of Psychiatry, vol. 54, no. 1, Jan. 2012, pp. 95–96. PubMed, https://doi.org/10.4103/0019-5545.94660.

Ayhan, Yavuz, et al. "Animal Models of Gene–Environment Interaction in Schizophrenia: A Dimensional Perspective." Progress in Neurobiology, vol. 136, Jan. 2016, pp. 1–27. DOI.org (Crossref), https://doi.org/10.1016/j.pneurobio.2015.10.002.

Bagot, Rosemary C., et al. "Variations in Postnatal Maternal Care and the Epigenetic Regulation of Metabotropic Glutamate Receptor 1 Expression and Hippocampal Function in the Rat." Proceedings of the National Academy of Sciences of the United States of America, 109 Suppl 2, no. Suppl 2, Oct. 2012, pp. 17200–07. PubMed, https://doi.org/10.1073/pnas.1204599109.

Bassett, Anne S., et al. "Copy Number Variations in Schizophrenia: Critical Review and New Perspectives on Concepts of Genetics and Disease." American Journal of Psychiatry, vol. 167, no. 8, Aug. 2010, pp. 899–914. DOI.org (Crossref), https://doi.org/10.1176/appi.ajp.2009.09071016.

Bassett, Anne S., and Eva W. C. Chow. "Schizophrenia and 22q11.2 Deletion Syndrome." Current Psychiatry Reports, vol. 10, no. 2, Apr. 2008, pp. 148–57. PubMed, https://doi.org/10.1007/s11920-008-0026-1.

Beards, Stephanie, et al. "Life Events and Psychosis: A Review and Meta-Analysis." Schizophrenia Bulletin, vol. 39, no. 4, July 2013, pp. 740–47. PubMed, https://doi.org/10.1093/schbul/sbt065.

Bodai, László, et al. "The Role of miR-137 in Neurodegenerative Disorders." International Journal of Molecular Sciences, vol. 25, no. 13, June 2024, p. 7229. DOI.org (Crossref), https://doi.org/10.3390/ijms25137229.

Boog, Georges. "Obstetrical Complications and Subsequent Schizophrenia in Adolescent and Young Adult Offsprings: Is There a Relationship?" European Journal of Obstetrics & Gynecology and Reproductive Biology, vol. 114, no. 2, June 2004, pp. 130–36. DOI.org (Crossref), https://doi.org/10.1016/j.ejogrb.2003.09.041.

Bosia, Marta, et al. "Schizophrenia, Cannabis Use and Catechol-O-Methyltransferase (COMT): Modeling the Interplay on Cognition." Progress in Neuro-Psychopharmacology & Biological Psychiatry, vol. 92, June 2019, pp. 363–68. PubMed, https://doi.org/10.1016/j.pnpbp.2019.02.009.

Bourque, F., et al. "A Meta-Analysis of the Risk for Psychotic Disorders among First- and Second-Generation Immigrants." Psychological Medicine, vol. 41, no. 5, May 2011, pp. 897–910. PubMed, https://doi.org/10.1017/S0033291710001406.

Brown, Alan S., et al. "Paternal Age and Risk of Schizophrenia in Adult Offspring." The American Journal of Psychiatry, vol. 159, no. 9, Sept. 2002, pp. 1528–33. PubMed, https://doi.org/10.1176/appi.ajp.159.9.1528.

Cantor-Graae, Elizabeth, and Carsten B. Pedersen. “Risk of Schizophrenia in Second-Generation Immigrants: A Danish Population-Based Cohort Study.” Psychological Medicine, vol. 37, no. 4, Apr. 2007, pp. 485–94. PubMed, https://doi.org/10.1017/S0033291706009652.

Cardno, A. G., and I. I. Gottesman. “Twin Studies of Schizophrenia: From Bow-and-Arrow Concordances to Star Wars Mx and Functional Genomics.” American Journal of Medical Genetics, vol. 97, no. 1, 2000, pp. 12–17.

Chen, Yun-zhou, et al. “Association of Histone Modification with the Development of Schizophrenia.” Biomedicine & Pharmacotherapy, vol. 175, June 2024, p. 116747. DOI.org (Crossref), https://doi.org/10.1016/j.biopha.2024.116747.

“Chlorpromazine Tablets.” Cleveland Clinic, https://my.clevelandclinic.org/health/drugs/18811-chlorpromazine-tablets.

Dalman, Christina, and Peter Allebeck. “Paternal Age and Schizophrenia: Further Support for an Association.” American Journal of Psychiatry, vol. 159, no. 9, Sept. 2002, pp. 1591–92. DOI.org (Crossref), https://doi.org/10.1176/appi.ajp.159.9.1591.

D’Arrigo, Terri, and Rachel Reiff Ellis. “Schizophrenia Symptoms.” WebMD, Mar. 2024, https://www.webmd.com/schizophrenia/schizophrenia-symptoms.

Dempster, Emma L., et al. “Disease-Associated Epigenetic Changes in Monozygotic Twins Discordant for Schizophrenia and Bipolar Disorder.” Human Molecular Genetics, vol. 20, no. 24, Dec. 2011, pp. 4786–96. PubMed, https://doi.org/10.1093/hmg/ddr416.

Doctoral School, Department of Neurosciences, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania, et al. “Genetic Polymorphism and Neuroanatomical Changes in Schizophrenia.” Romanian Journal of Morphology and Embryology, vol. 63, no. 2, Nov. 2022, pp. 307–22. DOI.org (Crossref), https://doi.org/10.47162/RJME.63.2.03.

Egan, Michael F., et al. “Effect of COMT $Val^{108/158}$ Met Genotype on Frontal Lobe Function and Risk for Schizophrenia.” Proceedings of the National Academy of Sciences, vol. 98, no. 12, June 2001, pp. 6917–22. DOI.org (Crossref), https://doi.org/10.1073/pnas.111134598.

Ermis, Aysegul, et al. “The Relationship between Catechol-O-Methyltransferase Gene Val158Met (COMT) Polymorphism and Premorbid Cannabis Use in Turkish Male Patients with Schizophrenia.” In Vivo [Athens, Greece], vol. 29, no. 1, 2015, pp. 129–32.

Forsingdal, A., et al. “15q13.3 Homozygous Knockout Mouse Model Display Epilepsy-, Autism- and Schizophrenia-Related Phenotypes.” Translational Psychiatry, vol. 6, no. 7, July 2016, p. e860. PubMed, https://doi.org/10.1038/tp.2016.125.

Grechuk, Katelyn, et al. “Cannabis, Schizophrenia Risk and Genetics: A Case Report of a Patient With Homozygous Valine Catechol-O-Methyltransferase Polymorphism.” Cureus, vol. 13, no. 6, June 2021, p. e15740. PubMed, https://doi.org/10.7759/cureus.15740.

Guimarães, Anna Luiza, et al. “Effects of Toxoplasma Gondii Infection on Cognition, Symptoms, and Response to Digital Cognitive Training in Schizophrenia.” Schizophrenia [Heidelberg, Germany], vol. 8, no. 1, Nov. 2022, p. 104. PubMed, https://doi.org/10.1038/s41537-022-00292-2.

Guo, Xinpeng, et al. “A Review on the Relationship between the Distal 1q21.1 Microdeletion and Schizophrenia.” Frontiers in Genetics, vol. 16, 2025, p. 1612654. PubMed, https://doi.org/10.3389/fgene.2025.1612654.

Harrison, G., D. Gunnell, et al. “Association between Schizophrenia and Social Inequality at Birth: Case-Control Study.” The British Journal of Psychiatry: The Journal of Mental Science, vol. 179, Oct. 2001, pp. 346–50. PubMed, https://doi.org/10.1192/bjp.179.4.346.

Harrison, G., C. Glazebrook, et al. “Increased Incidence of Psychotic Disorders in Migrants from the Caribbean to the United Kingdom.” Psychological Medicine, vol. 27, no. 4, July 1997, pp. 799–806. PubMed, https://doi.org/10.1017/s0033291796004643.

Harrison, Pj, et al. “Review: The Group II Metabotropic Glutamate Receptor 3 (mGluR3, mGlu3, GRM3): Expression, Function and Involvement in Schizophrenia.” Journal of Psychopharmacology, vol. 22, no. 3, May 2008, pp. 308–22. DOI.org (Crossref), https://doi.org/10.1177/0269881108089818.

Hilker, Rikke, et al. “Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register.” Biological Psychiatry, vol. 83, no. 6, Mar. 2018, pp. 492–98. DOI.org (Crossref), https://doi.org/10.1016/j.biopsych.2017.08.017.

Joseph, Jay. “The ‘Schizophrenia Is 80% Heritable’ Fallacy: Deconstructing a Twin Study Meta-Analysis.” Review of General Psychology, vol. 30, no. 2, June 2026, pp. 212–29. DOI.org (Crossref), https://doi.org/10.1177/10892680261428409.

Kendler, K. S. “Overview: A Current Perspective on Twin Studies of Schizophrenia.” The American Journal of Psychiatry, vol. 140, no. 11, Nov. 1983, pp. 1413–25. PubMed, https://doi.org/10.1176/ajp.140.11.1413.

Khokhar, Jibran Y., et al. “The Link between Schizophrenia and Substance Use Disorder: A Unifying Hypothesis.” Schizophrenia Research, vol. 194, Apr. 2018, pp. 78–85. DOI.org (Crossref), https://doi.org/10.1016/j.schres.2017.04.016.

Lakna. “What Is the Difference Between SNP and CNV.” PEDIAA, https://pediaa.com/what-is-the-difference-between-snp-and-cnv/.

LaMarco, Nicole. “How Has Schizophrenia Treatment Evolved?” Medical News Today, June 2026, https://www.medicalnewstoday.com/articles/timeline-of-schizophrenia-treatments.

Liu, Yong-ping, et al. “Rs1625579 Polymorphism in the MIR137 Gene Is Associated with the Risk of Schizophrenia: Updated Meta-Analysis.” Neuroscience Letters, vol. 713, Nov. 2019, p. 134535. DOI.org (Crossref), https://doi.org/10.1016/j.neulet.2019.134535.

Lopez-Castroman, Jorge, et al. “Differences in Maternal and Paternal Age between Schizophrenia and Other Psychiatric Disorders.” Schizophrenia Research, vol. 116, nos. 2–3, Feb. 2010, pp. 184–90. DOI.org (Crossref), https://doi.org/10.1016/j.schres.2009.11.006.

Ma, Jia-Hui, et al. “Association on DISC1 SNPs with Schizophrenia Risk: A Meta-Analysis.” Psychiatry Research, vol. 270, Dec. 2018, pp. 306–09. DOI.org (Crossref), https://doi.org/10.1016/j.psychres.2018.09.056.

Ma, Jingsong, et al. “Association Between the COMT Val158Met Polymorphism and Antipsychotic Efficacy in Schizophrenia: An Updated Meta-Analysis.” Current Neuropharmacology, vol. 19, no. 10, Oct. 2021, pp. 1780–90. DOI.org (Crossref), https://doi.org/10.2174/1570159X18666201023154049.

Magwai, Thabo, et al. “DNA Methylation and Schizophrenia: Current Literature and Future Perspective.” Cells, vol. 10, no. 11, Oct. 2021, p. 2890. PubMed, https://doi.org/10.3390/cells10112890.

Malaspina, Dolores, et al. “Advancing Paternal Age and the Risk of Schizophrenia.” Archives of General Psychiatry, vol. 58, no. 4, Apr. 2001, p. 361. DOI.org (Crossref), https://doi.org/10.1001/archpsyc.58.4.361.

Maric, Nadja P., et al. “Improving Current Treatments for Schizophrenia.” Drug Development Research, vol. 77, no. 7, Nov. 2016, pp. 357–67. DOI.org (Crossref), https://doi.org/10.1002/ddr.21337.

Mathew, R. J., et al. “Catechol-O-Methyltransferase and Catecholamines in Anxiety and Relaxation.” Psychiatry Research, vol. 3, no. 1, Sept. 1980, pp. 85–91. PubMed, https://doi.org/10.1016/0165-1781(80)90050-5.

McLoughlin, Benjamin C., et al. “Cannabis and Schizophrenia.” Cochrane Database of Systematic Reviews, edited by Cochrane Schizophrenia Group, vol. 2014, no. 10, Oct. 2014. DOI.org (Crossref), https://doi.org/10.1002/14651858.CD004837.pub3.

Modinos, Gemma, et al. “Molecular Genetic Gene–Environment Studies Using Candidate Genes in Schizophrenia: A Systematic Review.” Schizophrenia Research, vol. 150, nos. 2–3, Nov. 2013, pp. 356–65. DOI.org (Crossref), https://doi.org/10.1016/j.schres.2013.09.010.

Morgan, Craig, and Helen Fisher. “Environment and Schizophrenia: Environmental Factors in Schizophrenia: Childhood Trauma--a Critical Review.” Schizophrenia Bulletin, vol. 33, no. 1, Jan. 2007, pp. 3–10. PubMed, https://doi.org/10.1093/schbul/sbl053.

Owen, Michael J., et al. “Schizophrenia.” Lancet [London, England], vol. 388, no. 10039, July 2016, pp. 86–97. PubMed, https://doi.org/10.1016/S0140-6736(15)01121-6.

Patel, Krishna R., et al. “Schizophrenia: Overview and Treatment Options.” P & T: A Peer-Reviewed Journal for Formulary Management, vol. 39, no. 9, Sept. 2014, pp. 638–45.

Paylor, R. “Mice Deleted for the DiGeorge/Velocardiofacial Syndrome Region Show Abnormal Sensorimotor Gating and Learning and Memory Impairments.” Human Molecular Genetics, vol. 10, no. 23, Nov. 2001, pp. 2645–50. DOI.org (Crossref), https://doi.org/10.1093/hmg/10.23.2645.

Pearson, Nathan T., and James H. Berry. “Cannabis and Psychosis Through the Lens of DSM-5.” International Journal of Environmental Research and Public Health, vol. 16, no. 21, Oct. 2019, p. 4149. DOI.org (Crossref), https://doi.org/10.3390/ijerph16214149.

Pu, Xingfu, and Xiao Xiao. “No Evidence of an Association between MIR137 Rs1625579 and Schizophrenia in Asians: A Meta-Analysis in 30 843 Individuals.” Psychiatric Genetics,

vol. 26, no. 5, Oct. 2016, pp. 203–10. DOI.org (Crossref), https://doi.org/10.1097/YPG.0000000000000136.

Qin, Xianzheng, et al. “22q11.2 Deletion Syndrome and Schizophrenia.” Acta Biochimica Et Biophysica Sinica, vol. 52, no. 11, Dec. 2020, pp. 1181–90. PubMed, https://doi.org/10.1093/abbs/gmaa113.

Rees, Elliott, et al. “Genetics of Schizophrenia.” Current Opinion in Behavioral Sciences, vol. 2, Apr. 2015, pp. 8–14. DOI.org (Crossref), https://doi.org/10.1016/j.cobeha.2014.07.001.

Réthelyi, János M., et al. “Genes and Environments in Schizophrenia: The Different Pieces of a Manifold Puzzle.” Neuroscience & Biobehavioral Reviews, vol. 37, no. 10, Dec. 2013, pp. 2424–37. DOI.org (Crossref), https://doi.org/10.1016/j.neubiorev.2013.04.010.

Robinson, Natassia, and Sarah E. Bergen. “Environmental Risk Factors for Schizophrenia and Bipolar Disorder and Their Relationship to Genetic Risk: Current Knowledge and Future Directions.” Frontiers in Genetics, vol. 12, 2021, p. 686666. PubMed, https://doi.org/10.3389/fgene.2021.686666.

Saini, S. M., et al. “Meta-Analysis Supports GWAS-Implicated Link between GRM3 and Schizophrenia Risk.” Translational Psychiatry, vol. 7, no. 8, Aug. 2017, pp. e1196–e1196. DOI.org (Crossref), https://doi.org/10.1038/tp.2017.172.

Schizophrenia. https://www.nimh.nih.gov/health/topics/schizophrenia.

Schizophrenia. Mayo Clinic, Sept. 2024, https://www.mayoclinic.org/diseases-conditions/schizophrenia/symptoms-causes/syc-20354443.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. “Biological Insights from 108 Schizophrenia-Associated Genetic Loci.” Nature, vol. 511, no. 7510, July 2014, pp. 421–27. DOI.org (Crossref), https://doi.org/10.1038/nature13595.

Shokouhifar, Alireza, et al. “DISC1 Gene Polymorphisms and the Risk of Schizophrenia in an Iranian Population: A Preliminary Study.” Journal of Cellular Biochemistry, vol. 120, no. 2, Feb. 2019, pp. 1588–97. PubMed, https://doi.org/10.1002/jcb.27427.

Shorter, Kimberly R., and Brooke H. Miller. “Epigenetic Mechanisms in Schizophrenia.” Progress in Biophysics and Molecular Biology, vol. 118, nos. 1–2, July 2015, pp. 1–7. PubMed, https://doi.org/10.1016/j.pbiomolbio.2015.04.008.

Sipos, Attila, et al. “Paternal Age and Schizophrenia: A Population Based Cohort Study.” BMJ, vol. 329, no. 7474, Nov. 2004, p. 1070. DOI.org (Crossref), https://doi.org/10.1136/bmj.38243.672396.55.

Stępnicki, Piotr, et al. “Current Concepts and Treatments of Schizophrenia.” Molecules [Basel, Switzerland], vol. 23, no. 8, Aug. 2018, p. 2087. PubMed, https://doi.org/10.3390/molecules23082087.

Stilo, Simona A., and Robin M. Murray. “Non-Genetic Factors in Schizophrenia.” Current Psychiatry Reports, vol. 21, no. 10, Oct. 2019, p. 100. DOI.org (Crossref), https://doi.org/10.1007/s11920-019-1091-3.

Sullivan, Patrick F., et al. “Schizophrenia as a Complex Trait: Evidence From a Meta-Analysis of Twin Studies.” Archives of General Psychiatry, vol. 60, no. 12, Dec. 2003, p. 1187. DOI.org (Crossref), https://doi.org/10.1001/archpsyc.60.12.1187.

Swofford, Cheryl D., et al. “Double Jeopardy: Schizophrenia and Substance Use.” The American Journal of Drug and Alcohol Abuse, vol. 26, no. 3, Jan. 2000, pp. 343–53. DOI.org (Crossref), https://doi.org/10.1081/ADA-100100248.

Szecówka, Kamila, et al. “Copy Number Variations and Schizophrenia.” Molecular Neurobiology, vol. 60, no. 4, Apr. 2023, pp. 1854–64. PubMed, https://doi.org/10.1007/s12035-022-03185-8.

Thomas, Elizabeth A. “Histone Posttranslational Modifications in Schizophrenia.” Advances in Experimental Medicine and Biology, vol. 978, 2017, pp. 237–54. PubMed, https://doi.org/10.1007/978-3-319-53889-1_13.

Van, Lily, et al. “Update on the 22q11.2 Deletion Syndrome and Its Relevance to Schizophrenia.” Current Opinion in Psychiatry, vol. 30, no. 3, May 2017, pp. 191–96. DOI.org (Crossref), https://doi.org/10.1097/YCO.0000000000000324.

Van Winkel, R., et al. “Psychosocial Stress and Psychosis. A Review of the Neurobiological Mechanisms and the Evidence for Gene-Stress Interaction.” Schizophrenia Bulletin, vol. 34, no. 6, Aug. 2008, pp. 1095–105. DOI.org (Crossref), https://doi.org/10.1093/schbul/sbn101.

Van Winkel, Ruud, et al. “REVIEW: Genome-Wide Findings in Schizophrenia and the Role of Gene-Environment Interplay.” CNS Neuroscience & Therapeutics, vol. 16, no. 5, Oct. 2010, pp. e185-192. PubMed, https://doi.org/10.1111/j.1755-5949.2010.00155.x.

Varese, Filippo, et al. “Childhood Adversities Increase the Risk of Psychosis: A Meta-Analysis of Patient-Control, Prospective- and Cross-Sectional Cohort Studies.” Schizophrenia Bulletin, vol. 38, no. 4, July 2012, pp. 661–71. DOI.org (Crossref), https://doi.org/10.1093/schbul/sbs050.

Wahbeh, Marah H., and Dimitrios Avramopoulos. “Gene-Environment Interactions in Schizophrenia: A Literature Review.” Genes, vol. 12, no. 12, Nov. 2021, p. 1850. PubMed, https://doi.org/10.3390/genes12121850.

Wang, Alex W., et al. “Genome-Wide Association Study in Two Populations to Determine Genetic Variants Associated with Toxoplasma Gondii Infection and Relationship to Schizophrenia Risk.” Progress in Neuro-Psychopharmacology & Biological Psychiatry, vol. 92, June 2019, pp. 133–47. PubMed, https://doi.org/10.1016/j.pnpbp.2018.12.019.

Yang, Albert C., and Shih-Jen Tsai. “New Targets for Schizophrenia Treatment beyond the Dopamine Hypothesis.” International Journal of Molecular Sciences, vol. 18, no. 8, Aug. 2017, p. 1689. PubMed, https://doi.org/10.3390/ijms18081689.

Zammit, Stanley, et al. “Self Reported Cannabis Use as a Risk Factor for Schizophrenia in Swedish Conscripts of 1969: Historical Cohort Study.” BMJ, vol. 325, no. 7374, Nov. 2002, p. 1199. DOI.org (Crossref), https://doi.org/10.1136/bmj.325.7374.1199.

From Colonial Exploitation to Child Labor: The Enduring Cycle of Poverty in West Africa's Cocoa Industry By Veronica Wang

Imagine walking into a gourmet candy store and seeing numerous rows of colorful sweets and delicately hand-made chocolate with various flavors, labeled as "high quality" and "imported from Europe." According to Statista, Americans consume about 8.2 pounds of chocolate per year per person. In 2022, chocolate companies earned "$206 billion in revenue" in total. By 2030, the industry was predicted to "generate as much as $263 billion." However, cocoa farmers often earn far less than "$2.15 per day," which is way below the World Bank's poverty threshold (Corporate Accountability Lab). Contrary to how the chocolate industry is portrayed in the media, most cocoa farmers work under dangerous conditions and struggle to pay bills for daily necessities. Worse still, "an estimated 1.56 million" of the children in the Ivory Coast and Ghana work as child labor (International Cocoa Initiative). The continued exploitation of children in West Africa is rooted in two main factors: the long-lasting effects of European colonialism and large companies' pursuit of maximum profit.

History of the Cocoa Industry

Forced labor on cocoa farms has been a major problem since the 1500s. As chocolate became more popular in Europe, countries such as Spain and France started establishing cocoa plantations throughout their colonies in the Caribbean and South America. The majority of the workers were slaves brought across the Atlantic through the slave trade. In the late 1800s, the Portuguese introduced the first large-scale cocoa farms to the islands of São Tomé and Principe in West Africa. There, thousands of Africans were transported and forced to work under horrendous conditions similar to those in the age of slavery. Investigation by Dr. Joseph Burtt revealed that the average adult death rate was about 11% per year, whereas the average death rate in Europe at the same time was 2% to 3%. As demand for chocolate increased, colonial labor systems also started to spread across West Africa. On the Ivory Coast, both the European colonial government and local elites used forced labor to establish cocoa farms. For example, to increase profit, the French colonial government even encouraged migration from neighboring colonies for cheaper labor. Colonial governments designed the cocoa industry merely to maximize their own profits, with no consideration for the farmers' lives and well-being.

Nonetheless, being a cocoa worker was what most people had to do to provide for themselves and their families. As a result, even after forced labor was abolished in 1946 and countries gained their independence, the economic structure inevitably remained. The Ivory Coast has continued to expand its cocoa industry using forced labor and child labor, becoming the world's largest cocoa producer. To sum up, economic pressures and labor systems back from the colonial times continue to shape West Africa's present social and economic structure.

Cycle of Poverty

The remnants of colonialism continue to undermine communities and perpetuate the cycle of poverty. Since cocoa farming contributes to such a crucial portion of West African countries' economy, many people still continue to work on cocoa farms. However, the prices of cocoa offered by big corporations are often extremely low, forcing farmers to turn to cheap child labor. The farmers barely earn enough to sustain themselves, let alone paying workers a living wage. According to the Corporate Accountability Lab, farmers still earn incomes way below the World Bank's poverty line. Their inability to properly pay for workers makes them turn to their own children or trafficked children to help with harvesting the cocoa. (Corporate Accountability Lab). In an interview with a West African cocoa farmer he admitted that employing children was unfair but said he did so because he needed workers and earned too little from cocoa production. He recognized that children deserve an education and acknowledged that forcing them to work instead is a form of exploitation. Many of these farmers want their children to be educated but cannot afford schooling. For example, in his interview with Food is Power, Mr. Zongo, both a farmer and a parent, shared that he is forced to involve their children in the farm because they are not paid enough for cocoa. The pay for their labor remains low because the industry chooses to exploit the farmers for greater profits (Food is Power).

Many children are also illegally trafficked to cocoa farms. Some children are even sent by their relatives to traffickers or farm owners, often because their families do not realize the dangerous working conditions or that the children will not have access to an education (Food is Power). These children are often coerced into working through financial incentives and not properly informed of the working conditions or the type of work. Some of them are then kept as illegal slaves and would be punished for working slowly or trying to escape. In fact, trafficking is so common that in one village in Burkina Faso, almost every mother has had one child trafficked to cocoa farms.

Working and living conditions on the farms are also extremely concerning. The farm owners usually provide child laborers with the cheapest available food, including corn paste, cassava and bananas from nearby forests. In some extreme scenarios, the children live in small windowless buildings with no clean water or sanitary bathrooms, sleeping on wooden planks at night (Food is Power). They are frequently exposed to agricultural chemicals and subjected to severe beatings that leave permanent scars. Even though conditions are horrible, some children still choose to work on these farms to earn extra income that helps provide for their families. Many children in West Africa grow up in extreme poverty, causing them to start working at a young age to help provide for their families (Food is Power). Many child laborers are between the ages of 12 and 16, but children as young as 5 have been found working on cocoa farms. As long as poverty is still prevalent in Africa, many children will still choose to work on cocoa farms despite the horrendous conditions.

Because of long term poverty and exploitation, child labor has become the social norm in West Africa. One example is how many children choose to lie about their ages to work on these farms. Despite their wish to get educated, many of them do not have the resources to achieve their goals and end up staying on farms to help their families. This is an endless vicious cycle of

not having access to education and being forced into labor from a young age. Without schooling, these children are not able to build their knowledge and obtain jobs with decent wages, continuing to be exploited throughout their life. Eventually, when these people have children, they will pass down the idea that it is normal for children to work in farms early on. Over generations, this way of living has become the social norm in West Africa. According to an interview with an International Labor Organization worker, parents do not want to intentionally send their children to work on farms. Instead, it is the sociocultural norm that encourages child labor. Parents think that child labor was just a normal part of growing up (International Labor Organization).

Ultimately, this cycle continues today because the pursuit of profit motivates companies to maintain the existing system. Chocolate companies prioritize maximizing their profits instead of paying farmers fair wages. Due to public backlash, big companies like Nestle and Hershey vouched in 2001 to completely stop using cocoa harvested by children. However, they have not upheld that promise to this day. In his interview with the Washington Post, Antoine Fountain, the managing director of an organization focusing on human rights abuses in the cocoa industry, said the companies only do the bare minimum necessary to address public criticism and media attention (Washington Post). Companies are still able to maximize their profits because there are no laws or regulations requiring them to monitor their cocoa supply chains.

Efforts to Break the Vicious Cycle

Nevertheless, steps have been taken to try to break the cycle of exploitation. Efforts have been expanded to all levels of society in West Africa, including government and local levels. One major project that has helped reduce the use of child labor is the ACCEL Africa project led by the International Labor Organization (ILO). With the help from communities, this project aims to help children learn in school instead of working on farms. According to ILO, this approach makes sure that the governments, the employers, and the workers all work together to help children stop working farms. They provided various alternative solutions, such as offering children school supplies and other essential materials to stay in school (International Labor Organization). On the community level, one great achievement has been the establishment of the community child labor monitoring committees. Communities work together to develop their own community action plans, which has successfully removed a lot of children from the dangerous farms and enrolled them in classrooms. The community leaders have also been on advocacy visits to meet local political leaders. Based on their efforts, a new school was built in the critical area (International Labor Organization). Anecdotal evidence was provided as the proof for the project's effectiveness. In the story of Elizabeth and Divine, two young girls from Ondo State underwent significant transformation with the support of the project. The girls got enrolled in school and were removed from cocoa farms where they worked under hazardous conditions. Currently, both of them are faring well and have become entrepreneurs (International Labor Organization). Not only does this project actively remove children from farms and put them in schools, but it also raises awareness. Advocates educate parents about the dangers of cocoa farms

and the long term benefits of education. As time goes on, the normalization of child labor is slowly being broken down.

Other projects aiming to reduce child labor directly deposit money to farmers' bank accounts, which enable them to send their kids to school and earn a living wage. One project is the Living Income Differential project (LID), implemented in 2019 in Ghana and the Ivory Coast. It added $400-per-ton to the price of cocoa to help boost farmer incomes to a decent level. (Pulitzer Center). Similarly, local projects aimed towards helping communities were launched in areas without formal banking. One of the most effective projects was the Village Savings and Loans Associations (VSLAs), or Caisses Villageoises d'Épargne et de Crédit (CFREC or AVEC) in French.VSLAs help farmers by providing extra money through loans and grants to support the savings groups they already have. This gives farmers better access to financial resources when they need them (Pulitzer Center). In VSLAs, members meet regularly to contribute money to a savings fund. They can borrow from this fund at an interest rate agreed upon by the group. At the end of the year, the total savings and profits are divided among members in proportion to how much each person contributes (Pulitzer Center). Specifically, VSLAs often target women since they typically have fewer financial resources and opportunities compared to men in the same community. The project is able to address the problem of banking and gender inequality. According to the Pulitzer Center, by improving women's financial independence, VSLAs produce positive effects that benefit not only women but also their families and the wider community (Pulitzer Center). For example, grocery trader Bediassou Anne has been positively impacted by this project. She says that CFREC has greatly helped with her children's education. Before joining the program, she and her husband often had difficulty paying tuition, which forced their children to work on their farm. With additional income, she is now able to contribute to household expenses and afford her children's schooling (Pulitzer Center). Statistics have also shown the effectiveness of VSLAs. By the end of 2024, ICI stated that its Child Labour Monitoring and Remediation Systems (CLMRS) was able to cover "347,018 households." Because of these systems, 44% of the children once found in child labour were no longer working after two consecutive monitoring visits (Pulitzer Center).

Conclusion

Communities have been, and are still, deeply affected by the long history of colonialism, which greatly contributed to forced and child labor on cocoa farms. Communities are stuck in this vicious cycle of being undereducated and exploited for cheap labor. Farmers are in need of cheap child labor because they do not earn enough to afford workers. Children choose to work on cocoa farms because they want to help out their families. By working on farms instead of going to school, these children are stuck being exploited for their entire life. When they grow up and start families, they pass on the idea that child labor is acceptable, reinforcing the social norm. However, the ultimate reason behind the severe exploitation is chocolate companies trying to maximize their profits by underpaying the workers for raw cocoa. Although progress to break this cycle has been slow, significant strides have been made. For one, some companies have

helped West African communities by building schools and helping agricultural cooperatives. Projects like ACCEL and VSLA also aim to educate individuals to change harmful social norms, actively remove children from farms and into schools, and provide families with money. It is important to note, nevertheless, that breaking this vicious cycle is an ongoing uphill battle far from complete. Communities, governments, and companies must continuously work together to eradicate child labor from the cocoa industry.

Works Cited

Anti-Slavery International. The Cocoa Industry in West Africa: A History of Exploitation. Anti-Slavery International, 2004, https://www.antislavery.org/wp-content/uploads/2017/01/1_cocoa_report_2004.pdf.

Corporate Accountability Lab. West Africa Cocoa Report 2023. Corporate Accountability Lab, 2023, https://corpaccountabilitylab.org/west-africa-cocoa-report-2023.

Food Empowerment Project. "Slavery in the Chocolate Industry." Food Empowerment Project, https://foodispower.org/human-labor-slavery/slavery-in-the-chocolate-industry/.

International Cocoa Initiative. "Child Labour in Cocoa." International Cocoa Initiative, https://www.cocoainitiative.org/issues/child-labour-cocoa.

OPHI. Africa Multidimensional Poverty Peer Network (MPPN): B56 Africa Online. Oxford Poverty and Human Development Initiative, https://ophi.org.uk/sites/default/files/B56_Africa_online.pdf.

Pulitzer Center. "Cocoa, Child Labour and Côte d'Ivoire: Emerging Change." Pulitzer Center, https://pulitzercenter.org/stories/cocoa-child-labour-and-cote-divoire-emerging-change.

Statista. "Chocolate Industry." Statista, https://www.statista.com/topics/1638/chocolate-industry/.

The Dark Side of Chocolate. Uploaded by Miki Mistrati, YouTube, https://youtu.be/-13kIm03sqw.

"The Child Labor Behind Your Chocolate." The Washington Post, https://www.washingtonpost.com/graphics/2019/business/hershey-nestle-mars-chocolate-child-labor-west-africa.

Beyond Intuition: Exploring the Role of Probability and Statistics in Tennis
By Neil Reddy

Abstract

The modern game of tennis is increasingly shaped by data driven insights. While talent and physical prowess dominate on the court, the use of probability and statistics plays a growing role in analyzing player performance, predicting match outcomes, and developing winning strategies. This paper hypothesizes that probabilistic models and statistical tools significantly enhance decision making in coaching, match preparation, and fan engagement. Drawing from real world case studies and analytical frameworks, the research explores how data science is reshaping tennis from pre-match forecasting and point-level modeling to live match analytics and player performance evaluation.

The findings suggest that predictive models can reliably estimate match outcomes, identify high-leverage moments, and personalize strategies for players based on opponent tendencies and surface types. Real-time win probability updates, descriptive and advanced player metrics, and AI-powered simulations are increasingly central to elite performance management. Furthermore, the integration of analytics into fan experiences through platforms like IBM SlamTracker shows the growing democratization of these tools. While challenges such as data privacy and algorithmic transparency persist, evidence points toward a future where analytics are not only tools of optimization but also instruments of deeper understanding and engagement. Ultimately, this paper argues that the evolving role of probability and statistics is transforming how tennis is played, coached, and consumed.

Introduction

The use of statistics in tennis has grown manifold over the past few decades. Once limited to tracking basic metrics like aces, double faults, and unforced errors, the statistical toolkit now includes probabilistic forecasting, machine learning models, and real-time performance tracking. This evolution is reflective of a broader shift across the sporting world where analytics has become a core element of strategy, training, and fan engagement.

Tennis is a high variance sport- where a single point can alter momentum and matches can be decided by the slimmest of margins- prediction models are particularly useful. Unlike many team sports, tennis offers a clean, structured dataset: one-on-one competition, relatively controlled conditions, and well-defined scoring units, which makes it especially suitable for applying statistical models and probabilistic reasoning to uncover patterns, assess performance, and anticipate outcomes. This makes tennis a fertile ground for applying advanced analytics and deriving actionable insights. From a coaching perspective, data can reveal subtle weaknesses in an opponent's second-serve tendencies or uncover patterns in shot placement under pressure. For players, such insights inform tactical decisions and preparation. At the same time, fans now engage with tools that track real-time win probabilities, rally lengths, and momentum shifts- bringing them closer to the game than ever before, including sports betting.

As the line between data and decision-making becomes increasingly blurred, tennis is evolving into a sport driven as much by code as by competition. **This paper argues that probability and statistical modeling are not just enhancing tennis- they are transforming it in slow, but new and improved ways.** By studying the use of tools like Elo ratings, Markov chains, Bayesian forecasting, and AI-driven platforms, we aim to understand how analytics are shaping not just professional performance, but also public perception and strategic evolution within the sport.

The modern tennis court is no longer just a stage for athleticism and intuition- it's a data-rich environment where probability influences every serve, rally, and decision.

Methodology

This study adopts a **mixed-method approach** that blends **literature review**, **analytical review**, and **quantitative analysis**. The literature review explores the theoretical underpinnings of probabilistic modeling in sports, focusing on foundational methods such as Elo ratings, Markov chains, Bayesian inference, and machine learning classifiers. System reviews of platforms like IBM SlamTracker, Infosys Tennis Insights, and Tennis Abstract provide insight into how analytics are operationalized in elite tennis environments.

The analytical component involves applying key statistical tools to historical match datasets to demonstrate practical applications. These include modeling match simulations using Elo differentials, calculating point-level win probabilities with Markov processes, and updating in-match forecasts using Bayesian techniques. Descriptive statistics and advanced metrics are also used to evaluate player performance trends, surface effects, and head-to-head dynamics.

Additionally, real-world **case studies**- such as the use of analytics by Novak Djokovic's team and the role of IBM Watson at Wimbledon- illustrate how data is employed at the highest levels of competition. These case studies help validate theoretical claims by showing concrete examples of analytics influencing strategy and outcomes.

Together, these methods aim to provide a comprehensive overview of how probability and statistics are transforming tennis- from performance analysis to prediction and engagement- while also critically evaluating the limitations and ethical implications of such tools.

Probability in Match Prediction

Out of many various match prediction platforms, many use tools such as Glicko Rating, Markov Chains and more to help determine the most accurate result. This section explores Elo ratings and other probability tools and the role they play in accurate match forecasting.

I.Elo Ratings & Win Probability

The Elo rating system essentially assigns each player a numerical rating to represent their skill level, and this rating fluctuates based on match results(Jefferies, 2025(11). Elo ratings are primarily used to **predict win probabilities**, not actual match scores. The way the Elo rating is used to help determine the outcomes of tennis matches is by using the difference in Elo ratings

between two players to determine the **win probability** of a match. For example, if we were to take Rafael Nadal's Elo rating (1869) and Novak Djokovic's Elo rating (2160), we can use these numbers to estimate the outcome of the match. The formula used to calculate win probability is:

Win Probability for Player A = 1 / (1 + 10^((Rating of B - Rating of A) / 400))

Using this formula: **Win Probability for Nadal** = 1 / (1 + 10^((2160 - 1869)/400)) ≈ **0.24 or 24%** This means Nadal would have a 24% chance of beating Djokovic based on their Elo ratings(11).

Elo rating also has the benefit of being much more accurate than regular ATP or WTA rankings. According to an analysis from FiveThirtyEight, "Elo ratings change immediately after a match, whereas ATP and WTA rankings update weekly. Unlike ATP rankings, which reward points based on tournament level, Elo ratings account for the quality of the opponent"(Angelini et al, 2022). Elo ratings account for the quality of the opponent. ATP rankings also use a rolling 52-week window, meaning past results eventually drop off. Elo ratings, on the other hand, only reflect the latest match performances. Finally, because Elo considers opponent strength and match performance, it provides a more accurate predictor of future match outcomes(Angelini et al, 2022(14).

Another form of rating system that is used for statistical analysis in tennis is called the **Glicko Rating System**. Glicko is a method used to assess the skill level of tennis players, offering more accuracy than the Elo Rating System. Developed by Mark Glickman, this system not only gives a player a rating but also includes a **rating deviation (RD)**, which measures the confidence in that rating, the lower the RD, the more reliable the rating is(Glickman, 2022(10). Unlike Elo rating, the Glicko Rating is updated more significantly when they have fewer matches, allowing faster adjustments in rankings, especially for new tennis players or inactive players. Here is an example showing exactly how Glicko ratings are calculated, created and calculated by various sources including Delta Data, Mark Needham and the CRAN Glicko documentation.

Imagine two tennis players:

- **Player A**: Initial rating of 1500, Rating Deviation (RD) of 200, and volatility (σ) of 0.06.
- **Player B:** Initial rating of 1400, RD of 30, and σ of 0.06.

They play a match, and Player A wins.

Step 1: Convert to Glicko-2 Scale

Glicko-2 uses a transformed scale:μ = (rating - 1500) / 173.7178

- φ = RD / 173.7178

So, for Player A:

- μ = (1500 - 1500) / 173.7178 = 0
- φ = 200 / 173.7178 ≈ 1.15

For Player B:
$\mu = (1400 - 1500) / 173.7178 \approx -0.575$

- $\varphi = 30 / 173.7178 \approx 0.173$

Step 2: Compute Expected Score
The expected score (E) for Player A against Player B is calculated using the formula:
$g(\varphi) = 1 / \sqrt{(1 + 3\varphi^2 / \pi^2)}$

- $E = 1 / (1 + 10^{(-g(\varphi) * (\mu - \mu_{opponent}) / 400)})$

Applying these formulas, we find the expected outcome for Player A.
Step 3: Update Rating and RD
After the match, Player A's rating and RD are updated based on the actual outcome (a win, so s = 1) and the expected outcome (E). The formulas involve computing a variance (v), a delta (Δ), and then updating the volatility (σ), RD (φ), and rating (μ).
Finally, convert μ and φ back to the original rating scale:

- New rating = $\mu * 173.7178 + 1500$

- New RD = $\varphi * 173.7178$wrr

II. Point-Level Probabilities

Another tool used to find out the possible outcome of a match are Markov Chains. Markov Chains are very effective in net games such as tennis: “Finite Markov chain modelling thereby can be used as a descriptive tool, as well as to gain insight in the relationship between sports behavior and outcomes. The transition matrix can be used to display the frequency of game actions in relation to the respective game structure”(Rothe, 2023(6). At the most granular level, statistical models begin by calculating a player’s probability of winning a point on serve or return. For example, top male players win around 65–75% of their service points, while return win percentages tend to be much lower. These point-level probabilities serve as key inputs into Markov models.(Rothe, 2023(6)

A Markov Chain is a mathematical system that undergoes transitions from one state to another based on certain probabilities. In tennis, each point, game, or set can be modeled as a ‘state,’ with transitions representing the outcome of a point (win or loss). The Markov Chains greatly help with mid-point analysis, which might help statistical tools find quicker and better ways of predicting the winning percentage of a singular point, rather than an entire match.(Chan, 2021(12)

The main focus to use a markov chain is to be able to gradually find out the outcome of a match by using point to point statistics mainly with serving and returning. We can use the example of The Jannik Sinner and Carlos Alcaraz 2025 Roland Garros Final, which is debatingly of the best finals in the history of tennis. In the match, Alcaraz had a lower rate of first serves in, first serve points won, and second serve points won than Sinner. This statistic shows how Sinner has a good advantage in all three aspects of this example. By using the Markov Chains, we can conclude that Sinner has a higher percentage of winning service points than Alcaraz.(Rothe and

Lames, 2022(15)

Here is a step by step model that shows the simulation of all 5 sets of the Roland Garros 2025 Final between Carlos Alcaraz and Jannik Sinner:

Set-State Definitions:

- State 0 (S2–0): Sinner up two sets (after winning Sets 1 & 2 6-4, 7-6(7-4))
- State 1 (S2–1): Sinner up 2–1 (after Alcaraz wins Set 3)
- State 2 (S2–2): Sets tied at 2–2 (after Alcaraz takes Set 4)
- State 3 (A wins): Alcaraz wins match (after Set 5)
- State 4 (S wins): Sinner wins match (absorbing, but didn't happen)

Table 1

Transition Probabilities

From State	P(State 3)	P(State 4)	P(Next State)
0	0	1	Not used, match continued via Set 3
1	1	0	Base on match result: Alcaraz always won next
2	1	0	—

In reality, after State 0, Sinner could have closed in Set 3 but here, Alcaraz won three straight sets, so:

- $P(0 \rightarrow 1) = 1.0$
- $P(1 \rightarrow 2) = 1.0$
- $P(2 \rightarrow 3) = 1.0$

And:

- P(…) to State 4 = 0 (Sinner didn't win more)

Insights from the Chain

- After trailing 2–0, Alcaraz "forced" the chain through each successive state to an ultimate win.
- Because this is a single-match model, probabilities are 1.0 for the observed trajectory.
- With just one match, probabilities aren't probabilistic, they're deterministic. A full season model would be more helpful to estimate realistic chances.

Conclusion

This Markov chain exactly reproduced the sequence of the 2025 Sinner–Alcaraz RG final:

- Sinner leads 2–0 → Alcaraz takes Set 3 → Alcaraz takes Set 4 → Alcaraz clinches Set 5 (tiebreak)

The model is deterministic, but shows how set-by-set chain states and transitions can be used to describe a tennis final. To generalize, you'd fit transition probabilities using many match data sets. (Abdessatar Ati, Patrick Bouchet, Roukaya Ben Jeddou 2024(21), (Jorge Bapptista, 2020(22), Eric Kramer, 2024(23).

III. Pre-match Forecasting Tools

Pre-match forecasting tools are widely used to predict the outcome of a match before it even begins. These platforms analyze a variety of statistics, such as player ratings, court surface preferences, head-to-head history, and win-loss records. By aggregating this data, they generate informed predictions about potential match outcomes.(Marty Swant, 2024(2)

These forecasts are especially significant in the context of sports betting, which has become a popular activity worldwide. For bettors, fans, and analysts alike, pre-match forecasting tools offer a data-driven approach to anticipating match results and understanding the factors that may influence them. As such, they play an increasingly important role in both casual and professional evaluations of tennis matches.

The following diagram (figure 1)demonstrates how a common pre match forecasting tool determines the possible outcome of a match

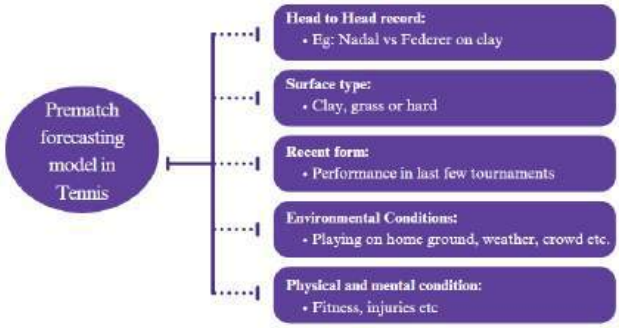


Figure 1

Prematch Forecasting Model in Tennis

Player Performance Analysis

I. Descriptive Statistics in Performance Evaluation

Understanding an opponent's statistics and playing style is crucial for effective match preparation. Descriptive statistics provide key insights into factors such as a player's strongest and weakest shots, preferred strategies, speed, and other technical attributes. These data points are used by players, coaches, and analysts to gain a competitive edge. A study published by the National Library of Medicine retrospectively analyzed 20,040 data points obtained from official ATP postseason sources. The researchers employed various analytical methods, including

inferential statistics and machine learning techniques, to examine factors like tournament participation, gameplay statistics, and prize money earnings. The study concluded that strategic participation in tournaments and consistent gameplay are essential for achieving higher rankings and maximizing financial returns. Additionally, the analysis of prize money distributions revealed significant disparities among players, emphasizing the importance of performance consistency.(Bozdech, Puda and Grasgruber, 2024) Overall, the study underscores the value of data-driven strategies in professional tennis, highlighting how specific gameplay and participation choices can impact a player's financial and ranking outcomes.

Figure 2 below demonstrates a heat map of first serve placement by Neil Reddy conducted during one of his practice sessions

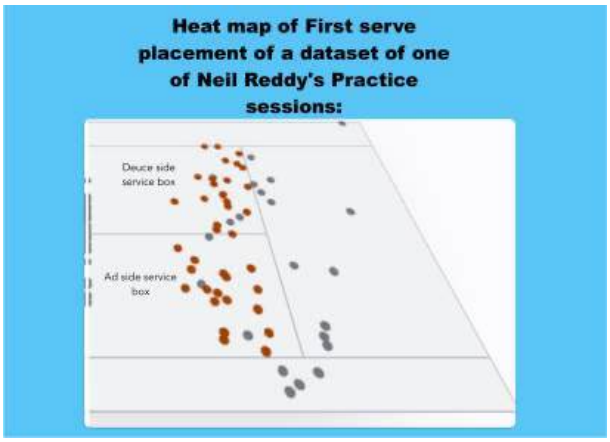


Figure 2
Heat Map of First Serve Placement of Neil Reddy's Practise session
Using this model, we can see that on the deuce side, Neil tends to serve more towards the T side half of the box, while on the ad side, he tends to serve more body and outside. Using models like these, we can find out where a player might tend to serve more, which is a stat that is very helpful for pre math preparation

II.Advanced Metrics

Advanced metrics in tennis analytics offer a deep understanding of player performance by going beyond regular statistics. One crucial metric is the expected hold/break percentage, which estimates the likelihood of a player holding serve or breaking their opponent based on past performance, court surface, and match conditions. Shot success probabilities are another essential tool, varying significantly by court type, such as higher forehand percentages on clay compared to hard court. Cluster analysis is used to define distinct player styles, grouping athletes with similar patterns in movement, shot selection, and point construction. These metrics help build strategy and player development. Vincent Auriau's GitHub project, *Tennis Match Predictions Using Machine Learning*, demonstrates how custom engineering and extraction can enhance model accuracy. His approach supports the creation of variables that help predict

models, improving projections for outcomes like break points or rally success based on surface and opponent tendencies.(Vincent Auriau, 2023)

Live Match Analytics

Live match analytics in tennis focuses on capturing real-time performance dynamics to better understand and predict match outcomes as they unfold. A key component is real-time stat tracking, which monitors metrics like serve accuracy, return depth, unforced errors, and rally length to detect momentum shifts. These changes indicate a mental or physical change in a player, influencing the flow of the match. Studies like *Analysis of Momentum in Tennis Matches Based on Ensemble Learning* emphasize the importance of data in identifying these transitions, allowing for tactical adjustments by coaches or analysts mid-match(Gong, Guo, Gao, 2024). Another vital tool is the evolution of in-match win probability. Using models like Markov chains, each point or game can be treated as a transition between defined states, producing updated probabilities after every point. *Markov Chain Modeling of Tennis* shows this, simulating match progressions based on probabilities tied to server strength, return ability, and historical tendencies. This provides an evolving win expectancy chart that can signal key turning points or upsets(Chan, Fearing, Fernandes and Kovalchik, 2021). Bayesian updating is another critical method for refining win predictions as new match data becomes available. As described in the Ghent University thesis *Predicting Tennis Matches Using Machine Learning*, Bayesian methods allow the incorporation of prior probabilities based on player statistics, surface type, or head-to-head history with live match input to continuously refine outcome forecasts(De Serrano, 2020). Similarly, the Stanford CS229 project, *Machine Learning for Professional Tennis Match Prediction and Betting*, shows the implementation of machine learning with Bayesian techniques to adjust predictions in response to real time performance changes, such as a player's drop in first-serve percentage or a sudden increase in double faults (Cornman, Spellman and Wright).

AI and Predictive Modeling in Tennis

Machine learning has revolutionized the way tennis matches are forecasted, offering more data-driven predictions rather than traditional statistics or expert opinion. These models go beyond win/loss records, incorporating nuanced features to reflect the complex dynamics of a match. One of the most critical aspects of predictive modeling is feature selection. Surface type—clay, grass, or hard court—directly influences player performance. For example, Rafael Nadal dominates on clay, while others excel on faster surfaces, such as hard courts.(De Serrano, 2020) Fatigue is another influential variable, often modeled through match duration or the number of consecutive playing days. Head to head records provide context on past encounters, while injury reports, when available, can explain unexpected performance drops. Instead of producing outputs such as win or lose, modern models provide probability distributions for each player's chances of winning. This approach allows analysts to understand confidence levels and adjust strategies accordingly. Bayesian methods and Monte Carlo simulations are often used to manage and interpret this uncertainty, offering more flexible and interpretable models. Ghent

University researchers achieved strong prediction accuracy using learning models, highlighting the importance of preprocessing and balanced datasets. Vincent Auriau's GitHub project demonstrates a practical application of scikit-learn based classifiers, complete with data scraping, cleaning, and model training. Lastly, a comparative study on ResearchGate evaluates Elo rating systems against ML methods, concluding that ML models when well done and supplied with robust features, can outperform traditional ranking systems.

Case Studies and Applications

Novak Djokovic's dominance in tennis is not only due to his athletic and mental strength, but also the strategic use of data by his team to prepare for opponents. His support staff which include coach Goran Ivanišević, a data analyst, fitness and mental performance coaches regularly use tools like Hawk-Eye, SAP Tennis Analytics, Dartfish, and SwingVision to gather and analyze match data. For example, before the 2021 Australian Open final against Daniil Medvedev, Djokovic's team studied Medvedev's return positions, rally tendencies, and weaknesses under pressure. They discovered Medvedev struggled in long rallies and often returned from deep behind the baseline. Based on this, Djokovic played aggressive tennis, used body serves to jam Medvedev, and forced errors by targeting the forehand. This preparation helped Djokovic win in straight sets. Typically, his team builds detailed pre match reports, creates video breakdowns, and simulates likely match scenarios based on opponent tendencies. By using detailed stats and analysis of his opponent, Djokovic and his team convert these stats into match specific strategies, proving that data is a vital part of modern tennis success.

IBM has played a big role in integrating advanced analytics into tennis, with its Watson platform becoming a big part of both Wimbledon and the US Open in recent years. At Wimbledon, IBM Watson is used to power SlamTracker, a tool that provides real time insights into serve percentages, rally length, break point conversion, and momentum shifts throughout matches. The system draws on historical player performance data and combines it with live match input to generate "Keys to the Match," predictive factors that highlight the strategies most likely to lead to victory. For example, Watson might indicate that Novak Djokovic needs to maintain a first-serve percentage above 65% or that Iga Świątek's chances of winning improve significantly when she wins rallies longer than nine shots. These insights not only give commentary but also enhance the fan experience by making data accessible through models and mobile apps. At the US Open, IBM has expanded Watson's role even further by introducing AI powered commentary and video highlight generation. Using language processing, the system can create real time written or spoken summaries of matches, automatically generating point by point recaps and match highlights within seconds. This is especially useful in a Grand Slam featuring hundreds of matches, where human editors cannot possibly keep up with demand. The AI commentary also helps bring out more of the fan experience, providing professional quality coverage even for matches on outside courts that received little to no media attention. IBM has also introduced Watsonx, an updated AI platform, at the US Open to improve transparency and adaptability of analytics, addressing some problems. However, these innovations have raised

concerns about accuracy and the accuracy of machine generated insights, with some critics suggesting that AI commentary may oversimplify or misrepresent parts of a match. Despite these limitations, IBM's partnership with Wimbledon and the US Open shows how AI is transforming tennis into a data driven sport, merging predictive modeling with fan engagement. By combining machine learning, natural language processing, and large scale data management, IBM Watson has made analytics a huge part of both performance monitoring and audience experience.

Discussion

The findings of this research shows that probability and statistical models have become a big part of modern tennis, not only in predicting outcomes but also in creating strategies, coaching, and fan engagement. Elo and Glicko ratings, for example, demonstrate how simple rating systems can outperform traditional ATP/WTA rankings by accounting for opponent quality and producing dynamic, match specific forecasts.

At the point level, Markov Chains reveal how small changes in serve and return percentages accumulate into important shifts in set and match probabilities. The 2025 Roland Garros final between Carlos Alcaraz and Jannik Sinner show the explanatory power of Markov models, showing how even these can highlight advantages or weaknesses. By integrating point-level probabilities with player-specific tendencies, Markov Chains provide coaches with insights on when and where players are most vulnerable.

The growth of machine learning and Bayesian forecasting adds another layer of information to the tennis world, bringing continuous updating of win probabilities as matches unfold. Real time analytics such as IBM SlamTracker and Infosys Tennis Insights combine descriptive metrics with predictive models, enhancing not only coaching decisions but also fan experience.

The application of these tools is perhaps most visible in professional player case studies. Novak Djokovic's team demonstrates how analytics can directly influence tactical choices, from targeting Medvedev's forehand to exploiting return positions. Such examples highlight the practical implications of statistical analysis, where probability models move beyond abstract predictions to inform concrete competitive strategies. They also validate the central hypothesis of this paper: that analytics are not merely supportive but transformative in shaping modern tennis.

Still, several limitations and challenges remain. Concerns about data privacy and algorithmic transparency are still in the air, particularly as AI-driven systems like IBM Watsonx become increasingly embedded in tennis commentary and analysis. Critics argue that these systems can appear more like branding exercises than genuine innovation, while inconsistencies in platforms such as Elo ratings reveal the need for continuous work. Furthermore, while predictive models can describe "what" is happening with great precision, they often fail to explain "why." As Andre Agassi notes, the interpretive burden still lies with players and coaches, reminding us that analytics must complement rather than replace human intuition.

In sum, the discussion highlights that probability and statistics are not peripheral to tennis, they are becoming the sport's central language. From predictive ratings to real-time

updates, from fan platforms to elite coaching, analytics are reshaping how tennis is played, experienced, and understood. The transformation is ongoing, and while challenges remain, the trajectory suggests that tennis is entering a new era where code and competition are the new focus.

Challenges and Ethical Considerations

There are of course many challenges and concerns though of what goes into statistical analysis of tennis matches, which might include restraints such as Data Privacy and consent, as well as algorithmic transparency. There have been many talks about data privacy in tennis statistics, including IBM's new Watsonx platform which combines AI and commentary which is likely to be used in the many coming US open matches: "Tennis fans shouldn't be surprised to see IBM's Watsonx platform bringing even more AI commentary to the 2024 US Open. It's been a partner with the tournament so long, it could be playing doubles"(Marty Swant, 2024). This new system is taking off quickly, and has caught the attention of many. Some, tend to look forward to this new and improved system, while some say that it is just a branding and marketing move from the Open. Another challenge can include the algorithmic transparency of these systems. Platforms such as IBM Slamtracker and Elo ratings in tennis have been analyzed by many, stating that they have many problems with it and it is becoming very inaccurate. Accordingly, IBM has created a new platform, while elo rating has become more and more secure with their calculations. Many problems will arise in the world of tennis and its statistical analysis, but it's important that we always try to fix it and solve its problems.

Future Outlook

The future of statistics in tennis is unpredictable, with the world constantly changing with AI. This has and will greatly impact the tennis world in the future as more and more new and improved systems and ratings will be brought up. It's important that all of these platforms and tools are brought to help though, and should be verified for any problems. Platforms such as elo rating and IBM have already incurred several problems, leading to the fix and improvement of these platforms. According to Agassi, using AI could help analyze what or how something happened, but leaves it up to the players and coaches to figure out why something happened. Although AI might provide players with "one hell of a roadmap" to better understand themselves(Marty Swant, 2024). This all just shows how quickly and efficiently the world of tennis is evolving through statistics and AI.

Conclusion

This paper has explored how probability and statistical analysis are not only helping with match predictions but reshaping how the sport is played, coached, and experienced. From pre-match forecasting using Elo and Glicko ratings to point level analysis with Markov Chains and real time win probabilities updated through Bayesian methods, analytics are now easily shown to be helpful in competitive success. Players and coaches use descriptive statistics

and advanced metrics to uncover strategic insights, optimize training, and develop opponent-specific game plans. Platforms like IBM SlamTracker and Tennis Abstract offer depth, providing in match adjustments and improved performance. Meanwhile, machine learning models, as shown by projects from Ghent University and Vincent Auriau, demonstrate that predictive accuracy improves significantly when AI is integrated with knowledge. However, this all is not without challenges. Concerns around data privacy, algorithmic transparency, and the reliance on automated systems remind us that human interpretation remains essential. As Agassi noted, data may tell us *what* happened, but understanding *why* still depends on the human mind. Looking ahead, the development of artificial intelligence and advanced analytics will continue to deepen, offering new possibilities for evolution and fan engagement. As tennis embraces its data driven future, probability and statistics will remain its most powerful allies on and off the court.

Works Cited

"Elo Ratings on Tennis Match Predictions." https://www.coresportsbetting.com/elo-ratings-on-tennis-match-predictions/

"Weighted Elo Rating for Tennis Match Predictions." https://www.sciencedirect.com/science/article/abs/pii/S037722172100324

"Glicko Rating System." https://www.glicko.net/glicko.html

Rothe, Florian, and Michael Lames. "Simulation of Tennis Behaviour Using Finite Markov Chains." https://www.sciencedirect.com/science/article/pii/S2405896322013672

"A Markov Process Approach to Untangling Intention Versus Execution in Tennis." arxiv.org

How IBM and the US Open are using WatsonX to create more AI-generated tennis content https://digiday.com/media/how-ibm-and-the-us-open-are-using-watson-create-more-ai-generated-tennis-content/

A detailed analysis of game statistics of professional tennis players: An inferential and machine learning approach https://pmc.ncbi.nlm.nih.gov/articles/PMC11537396/

Tennis Match Predictions Using Machine Learning https://github.com/VincentAuriau/Tennis-Prediction?tab=readme-ov-file

Analysis of Momentum in Tennis Matches Based on Ensemble Learning https://www.researchgate.net/publication/383105448_Analysis_of_Momentum_in_Tennis_Matches_Based_on_Ensemble_Learning

"Analysis of Momentum in Tennis Matches Based on Ensemble Learning."researchbank.swinburne.edu.au+4spiedigitallibrary.org+4pdfs.semanticscholar.org+4

Predicting Tennis Matches Using Machine Learning https://libstore.ugent.be/fulltxt/RUG01/002/945/727/RUG01-002945727_2021_0001_AC.pdf

Machine Learning for Professional Tennis Match Prediction and Betting https://cs229.stanford.edu/proj2017/final-reports/5242116.pdf

The Use and Interpretation of Popular Memes (Geng) on the Chinese Short-Video Platform Douyin: A Qualitative Sociolinguistic Study of Chinese Short-Video Culture

By Ki Kwong Kenrich Chen

Abstract

This study examines the use and interpretation of Geng, a type of popular meme or joke, on the Chinese short-video platform Douyin. Drawing on a qualitative multiple-case study of 20 videos and related user comments, the analysis identifies a spectrum of Geng types ranging from low-context sensory humor (e.g., the "Jet2 Holiday" audio) to high-context cultural performances (e.g., the "American bullying" trope). The findings show patterned manipulation of language and visual elements through "translationese," stylization, and multimodal resources. These recurring patterns can be interpreted as evidence of metapragmatic awareness. The discussion considers three possible social functions of these Geng practices: supporting shared recognition among users, framing uncomfortable content as playful and less realistic, and reframing everyday mishaps through humor. These functions illustrate how recurring Geng forms can support lightweight affiliation and playful participation in the sampled videos and comments.

Keywords: sociolinguistics; popular meme; Geng; short-video platform

1 Introduction

In recent years, Geng (lit. popular meme or joke) has become a common type of digital humor on Chinese short-video platforms. A Geng usually spreads as users repeat a feature from an earlier video and adapt it in other videos. To viewers unfamiliar with the relevant context, this kind of content may seem nonsensical or difficult to understand. However, its continued use suggests that many viewers are still able to recognize familiar parts of it even when the wording or scene changes.

This article has two main aims. The first is to examine how much background knowledge is involved in interpreting different examples of Geng. The second is to study how recognizable linguistic and visual features are reused and changed across the sampled videos and comments. The observations are based on the short-video platform Douyin, the Chinese version of TikTok. A study in this direction has significance for both academic theory and real-world practice. Theoretically, this analysis can echo and extend Verschueren's concept of metapragmatic awareness from a sociolinguistic approach (445). Practically, the insights generated from this study can contribute to a better understanding of how short-video users recognize and participate in shared forms of online humor.

2 Literature and Contexts

Existing studies have generally described internet memes as related versions of the same joke or format that spread through repetition and small changes. For example, Shifman suggests

that a meme is a collection of digital items that share elements of their content or form and circulate in different versions. Wiggins and Bowers describe each new version as a contribution to an ongoing online conversation. This kind of conversation follows an existing pattern and adapts that pattern to a new post or situation. Through a large-scale study of meme families, Segev et al. have also shown that a recurring phrase or familiar character can connect different versions to the same meme family. Together, these studies show that a meme can change considerably and still remain recognizable as part of the same family.

Other research has also shown that recognizing a meme is associated with what viewers have seen in earlier versions of the meme. For example, Dynel has shown that an image macro may be difficult to understand when viewed separately from the larger meme family it belongs to. Nissenbaum and Shifman argue that familiarity with meme conventions can become a form of "cultural capital" in online communities. Miltner's study of LOLCats has also pointed out that different populations may interpret and judge the same meme differently. Moreover, Jiang and Vásquez have shown that Chinese image memes can gain new meanings when the same type of image is combined with local cultural references. Together, these studies suggest that viewers often need earlier versions or cultural contexts to understand a new meme version.

Another topic of interest is the forms of online humor related to Geng. Park and Zhang have discussed the development of the meaning of the Chinese character Geng and shown that it has come to refer to online references that can be reused. De Seta's study of "biaoqing" (emoji) also describes how Chinese users modify visual materials and pass them on in daily online interaction. Guo shows how Chinese internet catchphrases can move beyond their original context and become part of wider online culture. Together, this work suggests that Chinese online humor often develops by carrying a familiar phrase or image into a new setting.
Another strand of research focuses on cases in which the wording itself is part of the joke. Verschueren uses the term "metapragmatic awareness" to describe people's awareness of how language is used in a particular situation. This idea is useful for meme research because many memes make a phrase funny by making the wording itself noticeable. Xuemei Chen's research on Chinese "dubbese" memes shows that translated-sounding Chinese can become a recognizable comic style on social media. Both studies suggest that the wording of online humor should also be seen as part of the joke itself.

Research on multimodal memes further examines the relation between words and images. Dancygier and Vandelanotte have argued that internet memes can be seen as combinations of verbal texts and visual materials, with the caption and the image interpreted together. Vásquez and Aslan's study of cat memes finds that humor often comes from the relationship between the caption and the image. Piata's research on classical art memes also shows that captions can change how viewers interpret old artwork. Their studies suggest that meme analysis should consider the caption and the image together.

Research on danmu (lit. bullet comments) and video comments have also examined viewer interaction around online videos. Yue Chen et al. have studied danmu-ku videos and argued that time-synced comments can make viewers feel as if they are watching a video with

others. Wu et al. have compared danmu with forum comments, showing that different comment formats support different forms of participation. Zhang and Cassany's study of the online video platform Bilibili's danmu also treats large numbers of anonymous comments as meaningful interaction around video content. These studies show that comments are already an important object in research on online video culture. The present study builds on these works to focus further on comments that reuse materials from the videos themselves.

Despite this body of seminal research, there are still two key research gaps that have not yet been covered. Specifically, Shifman, Wiggins and Bowers, and Segev et al. have explored the spread and changes of internet memes, while studies by de Seta, Guo, Jiang and Vásquez, and Chen have examined Chinese online humor through biaoqing (emoji), catchphrases, image memes, and dubbed memes. However, less attention has been given to how different examples of Chinese short-video Geng vary in the background knowledge they require, and to how comments continue a Geng by reusing material from the videos themselves. A comment may not simply repeat a phrase from the video but adapt the same speaking style or format to a new situation. In line with these two gaps, the present study asks two research questions. First, how does the background knowledge differ across the Geng discussed in this analysis? (RQ1) Second, how are recognizable linguistic and visual features reused and remade across the sampled videos and comments? (RQ2) Answering these two questions can contribute to our established knowledge about the circulation of popular memes in online social media contexts.

3 Methods and Materials

This study adopts a qualitative multiple-case-study approach using digital content analysis and multimodal analysis. The researcher drew on short videos on Douyin, the Chinese version of TikTok, and user comments on these videos. The author first conducted text analysis of the user comments. Following this, the author conducted multimodal analysis based on the screenshots of videos together with comment texts.

Specific to data collection, the researcher collected a primary dataset including 20 videos from Douyin. This dataset included the top 10 videos regarding the Geng of (American) bullying based on the number of "likes" received. This dataset also included the other top 10 videos regarding the Geng of Jet2 Holiday based on the number of "likes" received. The researcher also collected the text of top-ranked comments for each video on Douyin between July 2025 and October 2025. In the text analysis part, key attention was paid to Videos #1, #9, #11, and #12 and their comments.

Specific to data analysis, the researcher used a simple thematic analysis approach. The videos and selected comments were reviewed comparatively. The analysis focused on two main features: the background knowledge needed to understand each example and the linguistic and visual materials reused across them. These two features guided the analysis. The primary category was the type of Geng, including (American) bullying and Jet2 Holiday. The second category included three topics to be discussed in the following chapters, namely "the spectrum of

Geng's contexts," "metapragmatic awareness as the logic of patterned awkwardness," and "the social significance behind the nonsense."

4 Findings: The Spectrum of Geng's Contexts

To deconstruct "Geng," an analysis should not be restrained by its absurdity but should better frame it as a spectrum of information. This means that Gengs can be categorized by the amount of background information that a user, reader, or audience needs to understand them. Here, this section will focus on differences in background-knowledge demands.

At the far left, some Gengs require relatively little background knowledge to understand and use. These Gengs are usually understandable because much of the joke is clear from what viewers hear and see in the video. For example, the audio of "Jet2 Holiday" has a catchy and repetitive rhythm. The rhythmic song coupled with the thrilling commercial dubbing appears repeatedly across the sampled videos. Viewers do not need to know the full history of the original advertisement to notice the contrast created in many of these videos. In several sampled videos, users add another layer to the audio: they juxtapose the upbeat, cheerful audio against chaotic, "cringe" accidents. This creates a strong contrast between the sound and the image. When the video of a person engulfed in a massive torrent of water in an amusement park meets the "Jet2" jingle, the humorous effect comes from the mismatch between the cheerful audio and the accident shown in the video. Viewers can recognize this contrast without knowing the full development history of the Geng. Much of the humor is carried by the relationship between the sound and the image, therefore the Jet2 Holiday meme could be placed at the lower-context end of the spectrum. At the far right of the spectrum, we can move from pure literary symbols to complex performative systems that also include visual languages. The threshold for understanding them is high, with some often requiring readers to have in-depth cultural knowledge. The Geng of American bullying is a perfect example. Compared with the relatively easy-to-understand contrast in the Jet2 examples, American-bullying Geng relies on a more complex combination of linguistic styles and visual conventions. The Geng of American bullying stems from a trope in Hollywood teenage movies. In these movies, there is always a dorky newcomer at an American high school who is bullied by the captain of the football team in the locker room, often accompanied by a cheerleader (who is also the girlfriend of the bully) and a craven, chubby close friend on the side. On TikTok, this trope has shifted from its original theme of conflict between good and evil to cliché mimicking. In the Chinese adaptations in the sample, the trope becomes exaggerated live-action performances and low-fidelity digital cut-outs. Creators mimic the bully's cliched performances to an exaggerated extent, doing stereotypical actions such as pushing the nerd against the locker and grabbing the nerd's collar. Then, as it developed, creators included creative movements which established a sense of humor, such as swaggering or doing awkward dance moves. Recently, this trope developed into a whole new form. In such a form, creators started to use crude "cut-outs" to represent the characters (the head shot of a dorky, pale teenager with glasses representing the "nerd" archetype, and a head shot of Erling Haaland signifying the bully) and the props (a dark-red baseball bat and a dark-red

varsity jersey) instead of mimicking real people. These cut-outs slide rigidly over a background image of a locker room.

In addition, creators use "translationese," speaking Mandarin using the awkward, literal syntax of old Western movie dubs. Translationese occurs when English expressions are directly translated in a literal way without being adjusted for the correct Chinese syntax. Such translation hence results in a sense of crudeness. The degradation of characters, props, and language is used for humorous effect—when stiff language is combined with stiff action, the video becomes hysterical. Similarly, the cut-out style can be considered "visual translation." Just as these scripts break Mandarin syntax, these crude visuals break the consistency of a cartoon/movie. By doing so, these visuals use rigid translations (in the geometric sense of sliding an object across a plane) to show characters moving instead of walking naturally. Regardless of why creators selected this format, its recurring visual stiffness is similar to the translated-sounding dialogue and makes the parody recognizable across several sampled videos.

This spectrum suggests that these forms of Gen Z internet humor reflect patterned uses of linguistic and multimodal conventions. Whether using simple sensory jokes or complex cultural imitations, these conventions provide users with recognizable resources for participation and recognition.

5 Findings: Metapragmatic Awareness as the Logic of Patterned Awkwardness

The patterned uses of Geng can be interpreted as demonstrating metapragmatic awareness of spoken language and multimodal patterning of visual language. Linguist Jef Verschueren has argued that when people use language, they are not just putting words together but observing "how" and "why" they are put together. Jef Verschueren calls this ability not just to speak but to think about the act of speaking itself "metapragmatic awareness" (445). In the sampled videos, dialogues repeatedly differ from conventional syntax in highly patterned ways, making the form of the language itself noticeable.

It is worth noting that a person truly struggling with Chinese will make fragmented sentences using a limited vocabulary, while Gen Z users' "errors" are highly organized. The resulting wordplay depends on recognition of the original meaning and the selection of a different literal meaning. For example, the humor in the Chinese translation of "beats" in the Jet2 Holiday jingle depends on viewers recognizing both its idiomatic meaning, "is better than," and its literal meaning of "hitting someone." This patterned contrast suggests familiarity with the relevant language rules even as the videos depart from them.

In another form, this phenomenon of translationese also exists: Gen Z internet users will translate stereotypical expressions from the American bullying trope, such as "epic pool party" and "looks worse than my grandma's overbaked apple pie" (both of which are not common in Chinese expressions), to achieve the intended stiffness that generates humor. However, either the "synonym form" or the "expression form" follows a predictable pattern. Such consistency suggests a shared understanding of what conventional syntax looks like and how this shift can produce ironic humor. This way of speaking is what Slobe calls "stylization," in which

recognizable linguistic features are used to construct a social voice or persona (541). In the American-bullying parodies, this stylized voice also marks a visible distinction between the performers and the roles they portray. For example, a typical line from the American bullying parody is “oh, my old pal, look at what you’ve done!” In daily conversation, this word order and the term “old pal” would be uncommon in ordinary Mandarin conversation. The speaking of these cliched lines in a mocking tone clearly signals to the audience that the creators are only acting as a bully. This behavior exaggerates the difference between the performers and the roles they are playing, helping to frame the scene as parody rather than as a literal representation of bullying.

This control not only exists in the script but extends beyond it into the entire creation of the joke, as supported by what Li and Wang define as “vernacular creativity” through “multimodal resources.” Their research shows that internet users do not rely on typing as their only way of communication but quickly adopt the style of blending words with background music, stickers, and even glittering special effects to convey a nuanced meaning. This is evidenced by Video 9 (@Cccccc...), which shows the creative combination of glitching text (the Jet2 Holiday audio script) and remixed background music (Hold My Hand remixed in the style of a Chinese cheesy country song). This illustrates the controlled use of multimodal resources in the sampled videos. Without this mastery, Gen Z users cannot realize that words and beats can be synchronized, and thus the short videos will fail to create the sense of irony. This pattern can be interpreted as further evidence of metapragmatic awareness or multimodal reflexivity.
This manipulation is also visible in what this study calls “visual translationese”, which is a metaphorical term proposed by this study. In some comments and remixes, users extend the parody by introducing additional recognizable character types through celebrity headshots. Take the top-ranked comment from Video 11 (by @SUNNY L) as an example:

> “Jimmy, what happened to the injuries on your body?” (Video 11; translated by the author)

The commenter, in this context, is also the creator and uses the headshot of a cool and muscular actor, Dwayne Johnson, in a blazer to act as the nerd’s cool uncle. This is associated with dialogue such as “hey, Jimmy. Is that a bruise on your forehead?” This may resonate with many viewers, reminding them of a vague scene where the nerd’s FBI (Federal Bureau of Investigation) agent uncle notices the wound and punishes the bully. This resonance not only creates a connection between users but also results in a sense of humor. This example further illustrates how visual materials can be reused to extend the recurring character system of the Geng.

However, these communications are moot unless mutual understanding is established between the creator and the viewer. It is like a locked box built from awkward words and crude images. The creator secures the information in the box, but the box remains “empty” or “valueless” unless the audience has the specific cultural key to unlock or “decode” it. Over time,

a kind of creator called “Geng explainers” has come into being, analyzing the details of each Geng. Video 1 (by @leixingren) for “Jet2 Holiday” and Video 12 (by @jianghu baixiaosheng) for “American bullying” are great examples here. These videos establish the contexts of these Gengs, provide related background information, and even trace them back to their conceptual origins. The existence of these guides shows that Gen Z internet users have built such an organized and consistent digital vernacular that can be examined and taught like a formal language. The existence of such explainers provides further evidence that these Geng practices contain recurring conventions that can be described and taught to unfamiliar viewers. These explainers therefore act like locksmiths, helping unfamiliar viewers acquire the cultural key needed to understand the Geng.

6 Reflections: The Social Significance Behind the “Nonsense”

This essay now turns to reflections. Because this qualitative study analyzes videos and comments rather than users’ experiences, it cannot directly establish creators’ motives or viewers’ psychological responses. It can, however, support interpretive observations about the possible social and psychological functions of these Geng practices.

Here, the key question is: why do Gen Z internet users consistently use this kind of informal language (translationese), give awkward performances (American bullying trope), or juxtapose audio and videos in distinct tones (Jet2 Holiday meme)? This analysis considers three possible social functions of these Geng practices. The first is to support shared recognition among users. The second is to frame uncomfortable content as playful and less realistic. The third is to reframe everyday mishaps through humor.

First, Gengs can support shared recognition and lightweight affiliation among users. When living in a digital world where social connections are weaker than ever, internet users explore the alternative form of internet connections that can be easily acquired. Miller describes this form of connection as “phatic culture,” which means that staying in touch is more important than delivering messages. This connection is achieved in a virtual space created by Geng users. Due to the barrier of background information, this space is exclusive to those who understand the real meaning of Gengs, thereby creating a shared resonance that provides emotional value (companionship) for insiders. When commenters repeat a familiar line, speaking style, or character type from a video, they show that they recognize the same Geng, creating a form of “ambient affiliation” (Zappavigna). This instant understanding suggests that, even without in-depth conversation, users may experience a lightweight sense of belonging through shared participation in Geng. This sense of belonging is protected by the difficulty of coding and decoding the language of Geng: because these weird ways of speaking sound like nonsense to outsiders, the background knowledge required to recognize a Geng can mark a boundary between familiar and unfamiliar viewers.

Second, exaggerated language and artificial visual styles can frame uncomfortable content as playful and less realistic. Applying “nonsense” is a way for Gen Z internet users to create distance between themselves and the content. This may also help viewers build up shields

that protect them from feeling anxious or overwhelmed. A short video can show the bully scene in a locker room where the bully grabs the nerd by the collar and pushes him against the locker, but Gen Z users' distorted syntax and performance style can help the audience avoid the sense of pressure and uneasiness that may make them uncomfortable. In contrast, when audiences see the crude cut-out of Haaland sliding stiffly on a static background and speaking awkward phrases with weird gestures, they can be reminded that they are merely watching a parody of American bullying. The crudeness of the performance signals that the scene should be interpreted as parody rather than as a literal act of bullying. Bateson argues that play depends on metacommunicative cues that separate actions that look like aggression from actual aggression and establish a frame in which "this is play" (185, 195). In these videos, stiff translations, exaggerated gestures, and visibly artificial cut-out figures function as such cues. Slobe's account of stylization is also relevant: creators combine recognizable linguistic or physical features to build a parodic social persona (541). These features can reduce the scene's sense of reality and create distance from its threatening content. Although this study does not directly measure viewers' emotional responses, the awkwardness may make potentially stressful material easier to engage with through humor. Third, Gengs can reframe everyday mishaps through humor. In contemporary society that prefers efficiency and perfection, the rejection of seriousness by using Gengs makes it a way to take the weight off people's shoulders. The use of Geng as a "coping mechanism" (Akram et al.) shows that this form of absurdity is not only helping people avoid "damage" but also letting people reduce stress proactively. In "Jet2 Holiday," daily mishaps are paired up with cheerful music and are thus deprived of their seriousness. Consequently, the audience may stop feeling sorry for people in the mishap but start to laugh with them in harmless ways. This shared laughter can be described as "normalization" (Gardner et al.). If they see thousands of users laughing at the same "hiccup," they might be relieved, comforting themselves that their own "hiccups" are all merely a normal part of life. Indeed, this provides people with a sense of social support that may eliminate their loneliness to some extent. Through this form of reframing, an inconvenient or embarrassing event becomes material that can be shared as humor.

7 Conclusion

These Gengs are patterned manipulations of linguistic and multimodal conventions. Using Gengs may allow Gen Z internet users to quickly identify other users who share similar cultural values and follow the same hidden rules. This analysis presented above has created an original dataset with 20 videos and related users' comments. Then, the researcher has deconstructed the representative Gengs such as "American Bullying," "Jet2 Holiday," which range from simple symbols to complex performances.

The analysis suggests that these practices reshape conventional rules of communication on online platforms. They may create a shared interpretive space in which ordinary linguistic and visual conventions are temporarily altered. Within this space, users can display shared recognition and create distance between performers and the characters they portray. This distance may make potentially uncomfortable or stressful material easier to approach through humor.

However, this article also has its limitations, mainly because of the small sample size. Because this study analyzes videos and comments rather than users' self-reported experiences, its discussion of loneliness and stress should be understood as an interpretation of possible psychological functions rather than an assertion of users' internal states. Accordingly, potential quantitative research in the future will be helpful to verify the analysis results of this study.

Works Cited

Akram, Umair, et al. "Internet Memes Related to the COVID-19 Pandemic as a Potential Coping Mechanism for Anxiety." Scientific Reports, vol. 11, 2021, article 22305. https://doi.org/10.1038/s41598-021-00857-8.

Bateson, Gregory. "A Theory of Play and Fantasy." Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology, Jason Aronson, 1987, pp. 183–98.

Chen, Xuemei. "Dubbing and Reception: The Revival of Chinese Dubbese as Internet Memes on Social Media." The Journal of Specialised Translation, no. 44, 2025, pp. 85–104. https://doi.org/10.26034/cm.jostrans.2025.8470.

Chen, Yue, et al. "Watching a Movie Alone yet Together: Understanding Reasons for Watching Danmaku Videos." International Journal of Human–Computer Interaction, vol. 33, no. 9, 2017, pp. 731–43. https://doi.org/10.1080/10447318.2017.1282187.

Dancygier, Barbara, and Lieven Vandelanotte. "Internet Memes as Multimodal Constructions." Cognitive Linguistics, vol. 28, no. 3, 2017, pp. 565–98. https://doi.org/10.1515/cog-2017-0074.

de Seta, Gabriele. "Biaoqing: The Circulation of Emoticons, Emoji, Stickers, and Custom Images on Chinese Digital Media Platforms." First Monday, vol. 23, no. 9, 2018. https://doi.org/10.5210/fm.v23i9.9391.

Dynel, Marta. "'I Have Seen Image Macros!' Advice Animals Memes as Visual-Verbal Jokes." International Journal of Communication, vol. 10, 2016, pp. 660–88. https://ijoc.org/index.php/ijoc/article/view/4101.

Gardner, Kathryn J., et al. "Humor Styles Influence the Perception of Depression-Related Internet Memes in Depression." Humor, vol. 34, no. 4, 2021, pp. 497–517. https://doi.org/10.1515/humor-2021-0045.

Guo, Mengjun. "Playfulness, Parody, and Carnival: Catchphrases and Mood on the Chinese Internet from 2003 to 2015." Communication and the Public, vol. 3, no. 2, 2018, pp. 134–50. https://doi.org/10.1177/2057047318770467.

Jiang, Yaqian, and Camilla Vásquez. "Exploring Local Meaning-Making Resources: A Case Study of a Popular Chinese Internet Meme (Biaoqingbao)." Internet Pragmatics, vol. 3, no. 2, 2020, pp. 260–82. https://doi.org/10.1075/ip.00042.jia.

Li, Ziqi, and Li Wang. "Investigating Translanguaging Strategies and Online Self-Presentation through Internet Slang on Douyin (Chinese TikTok)." Applied Linguistics Review, vol. 15, no. 6, 2024, pp. 2823–55. https://doi.org/10.1515/applirev-2023-0094.

Miller, Vincent. "New Media, Networking and Phatic Culture." Convergence: The International Journal of Research into New Media Technologies, vol. 14, no. 4, 2008, pp. 387–400. https://doi.org/10.1177/1354856508094659.

Miltner, Kate M. "'There's No Place for Lulz on LOLCats': The Role of Genre, Gender, and Group Identity in the Interpretation and Enjoyment of an Internet Meme." First Monday, vol. 19, no. 8, 2014. https://doi.org/10.5210/fm.v19i8.5391.

Nissenbaum, Asaf, and Limor Shifman. “Internet Memes as Contested Cultural Capital: The Case of 4chan’s /b/ Board.” New Media & Society, vol. 19, no. 4, 2017, pp. 483–501. https://doi.org/10.1177/1461444815609313.

Park, Heungsoo, and Liang Zhang. “A New Study about the Meaning of ‘梗.’” The Journal of Chinese Character Studies, no. 27, 2020, pp. 23–44. https://doi.org/10.14772/CSCCK.2020.27.23.

Piata, Anna. “Stylistic Humor across Modalities: The Case of Classical Art Memes.” Internet Pragmatics, vol. 3, no. 2, 2020, pp. 174–201. https://doi.org/10.1075/ip.00031.pia.

Segev, Elad, et al. “Families and Networks of Internet Memes: The Relationship between Cohesiveness, Uniqueness, and Quiddity Concreteness.” Journal of Computer-Mediated Communication, vol. 20, no. 4, 2015, pp. 417–33. https://doi.org/10.1111/jcc4.12120.

Shifman, Limor. “Memes in a Digital World: Reconciling with a Conceptual Troublemaker.” Journal of Computer-Mediated Communication, vol. 18, no. 3, 2013, pp. 362–77. https://doi.org/10.1111/jcc4.12013.

Slobe, Tyanna. “Style, Stance, and Social Meaning in Mock White Girl.” Language in Society, vol. 47, no. 4, 2018, pp. 541–67. https://doi.org/10.1017/S004740451800060X.

Vásquez, Camilla, and Erhan Aslan. “‘Cats Be Outside, How about Meow’: Multimodal Humor and Creativity in an Internet Meme.” Journal of Pragmatics, vol. 171, 2021, pp. 101–17. https://doi.org/10.1016/j.pragma.2020.10.006.

Verschueren, Jef. “Notes on the Role of Metapragmatic Awareness in Language Use.” Pragmatics, vol. 10, no. 4, 2000, pp. 439–56. https://doi.org/10.1075/prag.10.4.02ver.

Wiggins, Bradley E., and G. Bret Bowers. “Memes as Genre: A Structurational Analysis of the Memescape.” New Media & Society, vol. 17, no. 11, 2015, pp. 1886–1906. https://doi.org/10.1177/1461444814535194.

Wu, Qunfang, et al. “Danmaku vs. Forum Comments: Understanding User Participation and Knowledge Sharing in Online Videos.” Proceedings of the 2018 ACM Conference on Supporting Groupwork, ACM, 2018, pp. 209–18. https://doi.org/10.1145/3148330.3148344.

Zappavigna, Michele. “Ambient Affiliation: A Linguistic Perspective on Twitter.” New Media & Society, vol. 13, no. 5, 2011, pp. 788–806. https://doi.org/10.1177/1461444810385097.

Zhang, Leticia-Tian, and Daniel Cassany. “Making Sense of Danmu: Coherence in Massive Anonymous Chats on Bilibili.com.” Discourse Studies, vol. 22, no. 4, 2020, pp. 483–502. https://doi.org/10.1177/1461445620940051.

The Role of Environmental Stressors on the Increased Risk of Polycystic Ovary Syndrome
By Niyanthri Jayakanthan

Abstract

Polycystic ovary syndrome (PCOS) is a common hormonal disorder affecting the ovaries in women during reproductive years. Although PCOS impacts approximately 1 in 10 women globally, up to 70% of cases go undiagnosed due to symptoms being frequently dismissed. Notably, a leading and often undiagnosed cause contributing to the prevalence of PCOS is stress: the body's physical and psychological response to challenges, demands, or threats, often created by a "flight-or-fight" response. The Office of the Surgeon General of the United States, a federal entity specializing in information on population health trends notes, "though certain forms of manageable, short term challenges can boost performance and motivation in day-to day life, chronic stress and cumulative biologic burden can contribute to worsened health outcomes"(Office of the Surgeon General, 32). Much like any other reproductive health disorder, the "flight-or-fight" response triggered by everyday stress can worsen PCOS by disrupting hormonal balance and promoting systemic inflammation throughout the body. With the frequency that stress is encountered on a day-to-day basis, it raises the question: To what extent do environmental stressors encountered in daily life contribute to the prevalence of PCOS in women? This investigation examines the role environmental stressors play in eliciting symptoms of PCOS and explores the dangers of untreated external stress on the development and prevalence of PCOS by looking at both psychosocial and anthropogenic environmental stressors.

Historical Background

Historically speaking, a prevailing perspective has been that PCOS stems mainly from genetic inheritance. First researched in 1935 by scientists Irving F Stein, and Michael L. Leventhal, historical perspective focuses on the triad of symptoms (Zeng et al.) such as amenorrhea (absence of menstruation), polycystic ovaries, and hirsutism (excess hair), all which is said to have originated from a specific pathology of the ovary. This claim was later supported in 1968, by H.E. Cooper and colleagues on their study on PCOS. The study had found that mothers and sisters of the patient showed inheritance of the same traits which led to the conclusion that PCOS is heavily impacted by its genetic foundation (Unluturk et al.). This belief, however, was seriously challenged in the early 1990s when a major shift in focus occurred after the 1990 National Institute of Health conference on PCOS. Despite early genetic findings, researchers began questioning whether heredity alone could explain PCOS's complexity. This period marked a transition from viewing PCOS as a strictly reproductive genetic disorder to recognizing it as a complex, metabolic, and polygenic trait influenced by stress as it "can contribute to poorer health-related behaviors" (Office of the Surgeon General, 27). Recently, in a study conducted by Altalf and colleagues, specialized researchers in the field of psychology and health utilized the perceived stress scale (PSS-10) ranging from 0-40 to assess the correlation between external stress and PCOS. The scale indicated that women with PCOS encountered

moderate to chronic stress daily (14-40), with the mean scoring of 22.93 or higher. In contrast, women without the syndrome recorded significantly lower mean stress scores, averaging around 19.56 and lower (Altalf et al.), suggesting that while external stress is a universal experience, its intensity is higher with PCOS. The reason external stressors may influence PCOS so heavily may be due to its ability to spike cortisol levels, as Chen and colleagues, all research scientists specializing in biology, explains high cortisol levels can cause "insulin resistance, which increases cardiometabolic risks and psychological burden" for women with PCOS (Chen et al.). Even with growing evidence of external stress accounting for the prevalence of PCOS, many researchers continue to emphasize the genetic basis of the disorder established decades ago, overlooking the deep connection environmental stressors can have in influencing the severity of PCOS in women across the globe.

Psychosocial Environmental stressors

Despite substantiating evidence, medical researchers continue to rely heavily on the genetic architecture of the syndrome, however, psychosocial environmental stressors can contribute to a greater extent compared to genetic factors. Psychosocial stressors refer to external life situations or social circumstances that create intense, often chronic, stress and challenges. For instance, a major psychosocial stressor present in many lives is social isolation. Social isolation is frequently linked to negative health outcomes as "people who are less socially connected may have increased logic burden which can contribute to worsened health outcomes susceptibility and weaker immune responses" (Office of the Surgeon General 23). A study conducted by Kamil Abdulla, professor and scientist specializing in biotechnology, and colleagues demonstrated that "social isolation significantly reduced FSH, LH, estrogen, and progesterone levels while increasing cortisol ($p < 0.05$)"(Abdulla et al.). Experimental subjects in this study were chronically isolated in enclosed spaces for over two weeks, where many faced "altered hormones and suppression of implantation-related gene expression" (Abdulla et al. 1766). These findings are notably striking given their correspondence to the common claim that the body actively adapts to a hostile environment, and that PCOS-related endocrine dysfunction may be influenced by similar pathways. Reforms to improve effective social connection such as strengthening bonds through spontaneous interactions, or building community organizations could reduce the medical impact of psychosocial environmental stress on PCOS.

Nevertheless, researchers argue that emotionally demanding or intense social episodes and connections— such as adverse childhood experiences (ACE's) or toxic relationships—can sometimes intensify the severity of PCOS as well. In a study conducted by Akansha Misra, a Ph.D. in Gender, Women's and Sexuality Studies and colleagues, it was found that "97.6% of respondents with a prior diagnosis of PCOS, also reported the highest proportion of childhood neglect and trauma"(Misra et al). Additionally, the study found that participants who encountered ACEs had formed negative connections that were often demanding and stemmed from verbal abuse in their childhood. These conditions seemed to have worsened PCOS and showed "significant overlap" with symptoms of PCOS such as "menstrual dysfunction and/or excess

male-like hair growth"(Misra et al.). Building on this, Karen D. Lincoln, a professor of environmental health, observed that "persons who experienced negative interactions with their friends are found to have high levels of psychological distress" (Lincoln), suggesting that conflicts and rash experiences can trigger distress which can inflame and weaken the rest of the body. This demonstrates that emotionally taxing connections foster elevated stress responses, spiking cortisol and aggravating the immune system, in turn intensifying PCOS, differing from positive or neutral social bonds that can eliminate isolation. Collectively, both perspectives highlight that psychosocial stressors—whether stemming from social isolation or harmful social connection—operate through similar neural mechanisms and hormones, ultimately influencing the endocrine system malfunctions associated with PCOS.

Anthropogenic Environmental Stressors

While psychosocial stressors play a significant role in shaping hormonal imbalances associated with PCOS, anthropogenic stressors have appeared as another general contributor due to their ability to act as an endocrine disruptor (ED). Anthropogenic stressors refers to stressors derived from changes occurring in the physical environment due to human activity. As global industrialization expands, such stressors are rising. For instance, according to the Department of Toxic Substances Control, in the United States alone, there are 80,000 chemicals currently in use, with roughly 1500 new substances introduced every year. Regardless of expanding knowledge on their harmful effects, anthropogenic stressors such as plastics and chemicals are relied on heavily by the population("Chemicals of Emerging Concern"). In a study conducted by Sana Ullah, a professional in endocrinology, and colleagues notes that bisphenol a (BPA), and industrial chemical seen in plastic was "reducing the growth of the follicles, decreasing the level of anti-Mullerian hormone [AMH] (99), estradiol, and causing an irregular estrous cycle and abnormal folliculogenesis" (Ullah). Additionally Gao et al., specializing in reproductive research, further found that BPA can activate estrogen receptors and inflammation pathways, promoting the progression of hormone-sensitive cancers (Gao et al.). These findings are particularly alarming as AMH is used to support ovarian development, and decreased levels of the hormone indicates infertility. Efforts to eliminate such toxic stressors such as minimizing pollution, and adopting sustainable practices could reduce both the environmental and bodily impacts contributing to women. Such measures would improve hormone balance and protect long term health by preventing exacerbating PCOS features therefore reducing the risk of associated complications such as endometrial disease or cardiovascular disease. Hafezi, a scientific researcher on diseases and cancer, and Rahman, a professor with a focus on molecular genetics, explained that BPA can alter the production of different types of collagen around different cells. This alteration of collagen can modify hormones, primarily estrogen, and lead to the disruption of follicular development (Hafezi and Rahman). As a result, individuals exhibit rapid follicle loss, particularly when exposed to high levels of EDC's through prolonged contact. In daily settings, frequent contact with such chemicals can worsen PCOS symptoms by promoting follicle loss. Understanding how anthropogenic stressors can damage the female body

is crucial to recognizing its potential to produce life threatening conditions and worsen symptoms of PCOS. Everyday exposure to anthropogenic stressors such as BPA through items such as plastic containers, water bottles, and food packaging, can intensify symptoms of PCOS to a greater degree. Still however, advocates for such anthropogenic developments propose that they don't solely pose as harmful contributors to the environment, and are used for greater sustainability and convenience. For instance, a basis for this argument stems from the perception that plastics contribute to the health and safety of consumers in food and water packaging applications—all needed for convenience and survival. Additionally, proponents for plastic such as Andrady, a biomolecular engineer, argue that these stressors such as plastics are lightweight, easy to manufacture, and are installed in a range of diverse water control and distribution systems.Without them, many rural areas would lack access to essential resources (Andrady).

Nonetheless, this perspective fails to acknowledge the severity of encountering these chemicals on a daily basis through such uses. While plastic becomes increasingly normalized, their cumulative effects on health can induce PCOS, as an expert in reproductive science asserts "during plastic manufacturing, different chemical compounds are added, such as polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons, phthalates, organochlorine pesticides, and bisphenol A (BPA), which can leach and act as EDs interfering with hormone systems and cause reproductive disorders" (Gallo). Unknowingly, women stay continuously exposed to factors such as EDs that gradually disrupt their reproductive function, as they are inside many chemicals. Additionally, exposure to heavy metals such as mercury can compromise DNA protection in reproductive cells, leading to PCOS due to lack of proper molecular structure (Gallo). Endocrine disrupting chemicals seen in these human-made stressors contribute to rapid reproductive aging and follicle loss, both of which increase PCOS. Ultimately, although anthropogenic stressors are common, they can significantly impair reproductive function and contribute to the development of PCOS.

Conclusion

Medical facilities should make diagnosing PCOS through a multi-faceted approach a top priority. This can be executed through lifestyle changes, social awareness, and psychotherapy. For example, recent research strongly supports a multimodal, personalized lifestyle approach as a highly effective method to reduce environmental stressors and mitigate symptoms. In a personalized lifestyle, women can incorporate biodegradable materials and minimize contact with harmful chemicals for better hormone regulation. For example, data from a 2021 study published by the University of Rochester, found that older women who followed a lifestyle minimizing plastic, canned goods, and processed foods had significantly lower urinary concentrations of bisphenol A (BPA), demonstrating that altering certain lifestyle stressors can minimize PCOS symptoms by reducing exposure which improves hormonal regulation (University of Rochester Medicine). Also, to lower cortisol levels emerging associated with chronic stress, women have turned toward Cognitive Behavioral Therapy (CBT), to process and heal internally from painful memories. Not only does CBT build stronger healthy connections

through therapy, as Liu and colleagues, all researchers in the medical field explain, it can also promote a healthier environment where stress is incapable of striving. While changing behavioral patterns, pathways are formed differently, which can significantly lower inflammation markers that cause PCOS (Liu et al). While PCOS is unable to be cured, this comprehensive approach is able to change lifestyle factors that can reduce the significance and risk of gaining PCOS. However, it is important to acknowledge limitations, such as accessibility to safer products, and time it takes to recover from the exposure to stress. For example, psychology and psychiatry experts Antoni and Dhabhar, establish that prolonged stress can affect the body for several years even during treatment, establishing that environmental stressors can be challenging to overcome.

Although stress is inevitable, it can introduce significant risk to health. Environmental stressors—particularly those related to psychosocial and anthropogenic causes—often go unnoticed, even when affecting well-being severely. These stressors can act as endocrine disruptors and pollute the body, overall influencing reproductive hormones and spreading inflammation throughout ovaries. Emerging research in gynecology suggests that stress acts as a suppressant of the needed hormones for a healthy reproductive system, and can worsen conditions by decreasing regulation. The familiarity and frequency of such stressors can compound and further distort PCOS, making it more challenging to properly diagnose. To ensure that conditions of PCOS are accurately and promptly diagnosed, the medical community must acknowledge PCOS as a multisystem chronic condition and integrate multidisciplinary care. This approach could transpire through the implementation of specialized clinics, and raising awareness in communities. Recognizing environmental stressors as a prevailing fuel for PCOS is essential for the diagnosis and treatment of the syndrome for women throughout the world.

Works Cited

Abdulla, Alaa Kamil, et al. “Molecular Impact of Social Isolation on Implantation Gene Expression in Uterine Tissue: A Rabbit Model for Stress-Induced Fertility Disorders.” Journal of Applied Biotechnology Reports, vol. 12, no. 3, Sept. 2025, pp. 1766–75, https://doi.org/10.30491/jabr.2025.548230.1915. Accessed January 2026.

Andrady, Anthony L., and Mike A. Neal. “Applications and Societal Benefits of Plastics.” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, no. 1526, July 2009, pp. 1977–84, https://doi.org/10.1098/rstb.2008.0304. Accessed January 2026.

Altaf, Kishwar, et al. “Role of Stress in Onset of Polycystic Ovarian Syndrome (PCOS).” Journal of Health and Rehabilitation Research, vol. 3, no. 2, Dec. 2023, pp. 1122–27, https://doi.org/10.61919/jhrr.v3i2.276. Accessed January 2026.

"Chemicals of Emerging Concern." California Department of Toxic Substances Control,2021, dtsc.ca.gov/emerging-chemicals-of-concern/. Accessed April 2026.

Chen, Xianrui, et al. “Insulin Resistance in PCOS: Pathophysiological Mechanisms of Menstrual Dysfunction and Evidence-Based Treatment Strategies.” Biology of Reproduction, vol. 113, no. 6, 2025, pp. 1340-1354, https://doi.org/10.1093/biolre/ioaf197. Accessed February 2026.

Gallo, Alessandra. “Reprotoxic Impact of Environment, Diet, and Behavior.” International Journal of Environmental Research and Public Health, vol. 19, no. 3, Jan. 2022, p. 1303, https://doi.org/10.3390/ijerph19031303. Accessed January 2026.

Gao, Hui, et al. “Bisphenol A and Hormone-Associated Cancers.” Medicine, vol. 94, no. 1, Jan. 2015, p. e211, https://doi.org/10.1097/md.0000000000000211. Accessed January 2026.

Hafezi, Shirin A., and Wael M. Abdel-Rahman. “The Endocrine Disruptor Bisphenol a (BPA) Exerts a Wide Range of Effects in Carcinogenesis and Response to Therapy.” Current Molecular Pharmacology, vol. 12, no. 3, Mar. 2019, https://doi.org/10.2174/1874467212666190306164507. Accessed January 2026.

Lincoln, Karen D. “Social Support, Negative Social Interactions, and Psychological Well‐Being.” The Social Service Review, vol. 74, no. 2, 2000, pp. 231–252, https://doi.org/10.1086/514478. Accessed January 2026.

Liu, Yun-Zi, et al. “Inflammation: The Common Pathway of Stress-Related Diseases.” Frontiers in Human Neuroscience, vol. 11, no. 316, June 2017, https://doi.org/10.3389/fnhum.2017.00316. Accessed February 2026.

Misra, Akanksha, et al. “The Impact of Childhood Abuse and Neglect on the Development of Features of Polycystic Ovary Syndrome: A Pilot Study.” Women’s Health Reports, vol. 6, no. 1, Jan. 2025, pp. 412–420, https://doi.org/10.1089/whr.2024.0130. Accessed February 2026.

Office of the Surgeon General of the United States. “Chapter 2: How Social Connection Impacts Individual Health and Well-Being.” Our Epidemic of Loneliness and Isolation: The US Surgeon General’s Advisory on the Healing Effects of Social Connection and

Community. 2023, pp. 23-35. College Board AP Seminar Performance Task 2: Individual Research-Based Essay and Presentation Stimulus Materials. Accessed January 2026.

"Rochester Study Raises New Questions About Controversial Plastics Chemical." University of Rochester Medical Center, 27 Jan. 2009, www.urmc.rochester.edu/news/story/rochester-study-raises-new-questions-about-controversial-plastics-chemical .Accessed Apr. 2026.

Ullah, Sana, et al. “A Review of the Endocrine Disrupting Effects of Micro and Nano Plastic and Their Associated Chemicals in Mammals.” Frontiers in Endocrinology, vol. 13, Jan. 2023, https://doi.org/10.3389/fendo.2022.1084236. Accessed February 2026.

Unluturk, Ugur, et al. “The Genetic Basis of the Polycystic Ovary Syndrome: A Literature Review Including Discussion OfPPAR-γ.” PPAR Research, vol. 2007, 2007, pp. 1–23, https://doi.org/10.1155/2007/49109.

Zeng, Ling-Hui, et al. “Polycystic Ovary Syndrome: A Disorder of Reproductive Age, Its Pathogenesis, and a Discussion on the Emerging Role of Herbal Remedies.” Frontiers in Pharmacology, vol. 13, July 2022, p. 874914, https://doi.org/10.3389/fphar.2022.874914.

How Effective Are Convolutional Neural Networks in Integrating Visual and Structured Data for House Price Prediction By Grace Sehgal

Abstract

Convolutional neural networks (CNNs) have grown exponentially over the past few decades, especially in the fields of pattern recognition and image processing. They are among some of the most commonly used neural networks in machine learning due to their ability to handle visual data much better than others because they recognize the spatial relationships between pixels. Through convolutional layers, CNNs can automatically extract hierarchical features from images, allowing them to identify patterns that may not be easily captured through traditional methods. While CNNs were originally developed for computer vision tasks, recent research has explored their ability to integrate visual information with structured data for more complex prediction problems. One emerging application of this multimodal approach is real estate valuation, where property images can provide additional information beyond traditional numerical housing features. This literature review will examine how effective CNNs are at combining visual and structured data for house price prediction. It focuses on the methods researchers have developed, their reported results, and the challenges that remain.

Keywords: convolutional neural networks; multimodal learning; house price prediction; computer vision; deep learning

Introduction

Convolutional neural networks (CNNs) serve as systems that handle visual data such as images by processing the data through convolutional layers and filters (LeCun, Bengio, & Hinton, 2015). These networks have become a building block in the field of computer vision specifically in tasks like image classification and pattern recognition (LeCun, Bengio, & Hinton, 2015).

The range of functions a CNN can perform is wide due to the fact that they can effectively handle large datasets because they're skilled at detecting important features such as shapes and patterns (Krizhevsky, Sutskever, & Hinton, 2017). Their architecture consists of convolutional layers, pooling layers, and fully connected layers which allow the network to process grid-like data (LeCun, Bengio, & Hinton, 2015). Through this layered structure, CNNs are able to learn increasingly complex representations, beginning with basic visual patterns and progressing toward higher-level features that are useful for prediction tasks (He et al., 2016).

As CNNs have continued to improve, their applications have expanded far beyond traditional image classification. Today, they're used in fields such as healthcare, autonomous vehicles, facial recognition, environmental monitoring and finance, where extracting meaningful information from visual data is essential (LeCun, Bengio, & Hinton, 2015). Their ability to automatically learn complex visual features has made them one of the most influential deep learning models in modern artificial intelligence. The development of architectures such as

ResNet and Inception has further improved CNN performance by allowing models to capture more detailed and complex visual representations (He et al., 2016; Szegedy et al., 2015).

One area where CNNs have gained increasing attention is house price prediction. Accurately estimating the value of a property is important for homeowners, buyers, sellers, investors and real estate professionals because it influences purchasing decisions, market analysis and investment strategies. Traditionally, house prices have been predicted using structured information such as square footage, location, number of bedrooms and number of bathrooms (Geerts et al., 2023). However, these numerical features don't capture visual characteristics of a property, such as architectural style, curb appeal, renovation quality or overall condition, all of which can influence a home's market value. Traditional automated valuation models have historically relied on these structured characteristics but researchers have increasingly explored additional sources of information to improve prediction accuracy (Geerts et al., 2023).

To address this limitation, researchers have begun to incorporate property images into predictive models using convolutional neural networks (Nguyen et al., 2023). More recently, studies have explored multimodal approaches that combine both visual and structured data. By integrating these two sources of information, researchers aim to create models that capture a more complete representation of a property (Baltrušaitis et al., 2019). For example, image-based models can identify visual characteristics such as property appearance and condition while structured features provide measurable information such as size and location (Nguyen et al., 2023). Combining these complementary sources of information may allow models to produce more accurate property valuations than models relying on only one data type (Nguyen et al., 2023).

However, combining image features with numerical data presents several challenges. Images produce thousands of learned visual features while structured housing attributes consist of only a small number of numerical variables (Baltrušaitis et al., 2019). This imbalance makes it difficult for models to effectively learn from both data types without one dominating the other. Multimodal learning research has identified similar challenges involving feature representation, alignment and fusion strategies when combining different forms of data (Baltrušaitis et al., 2019). Therefore, developing effective methods for integrating CNN-generated visual features with structured information remains an important area of research.

As research in this area continues to grow, a variety of CNN architectures and feature fusion techniques have been proposed to address these challenges. While many studies report improvements through multimodal learning, the approaches differ in their datasets, network architectures, fusion methods and evaluation metrics (Nguyen et al., 2023; Geerts et al., 2023). Comparing these studies provides a better understanding of the strengths, limitations, and current state of CNN-based house price prediction.

This paper presents a literature review examining how convolutional neural networks have been used to integrate visual and structured data for house price prediction. By synthesizing recent research, this review explores common modeling approaches, compares their reported

performance, identifies recurring challenges, and discusses future directions for CNN-based multimodal learning in real estate valuation.

Methodology

This paper uses a literature review methodology to examine how effectively convolutional neural networks (CNNs) integrate visual and structured data for house price prediction. This study analyzes existing research to understand how researchers have approached multimodal house price prediction, the results these methods have produced and the challenges that still remain.

The studies included in this review were primarily found through Google Scholar using keywords such as *convolutional neural networks, house price prediction, deep learning, multimodal learning, real estate valuation, computer vision, property images* and *structured data*. Research published between 2015 and 2025 was prioritized to focus on recent advancements in deep learning while also including foundational studies that introduced important CNN architectures and multimodal learning techniques.

The selected studies were chosen based on their relevance to the research question: How effective are CNNs in integrating visual and structured data for house price prediction? Studies were included if they explored the use of CNNs for processing visual data, investigated ways to combine visual and structured information or provided important background on deep learning and multimodal models. Studies that focused only on traditional valuation methods or machine learning approaches without relevance to CNNs were excluded.

Each study was analyzed based on its research focus, dataset, model architecture, types of data used, feature integration methods, evaluation metrics and major findings. Particular attention was given to how researchers combined visual information from property images with structured housing features, including measurable property attributes and characteristics. By comparing these approaches, this review examines whether integrating multiple forms of data allows CNN-based models to create more accurate house price predictions.

The findings from these studies were then compared to identify common patterns, differences between approaches, and limitations within current research. The literature was organized into major themes, including CNN-based visual feature extraction, structured data approaches, multimodal data integration and current challenges. Through this analysis, this literature review evaluates the effectiveness of CNN-based multimodal models and explores how future research can continue improving house price prediction methods.

CNNs as Visual Feature Extractors

Convolutional neural networks have become one of the most effective approaches for extracting meaningful information from visual data. Unlike traditional methods that require manually identifying important image features, CNNs are able to automatically learn patterns directly from images through multiple layers of processing (LeCun et al., 2015). Each layer of a CNN captures different levels of information, with earlier layers identifying simple features such

as edges, textures and shapes, while deeper layers recognize more complex patterns and structures (LeCun et al., 2015). This ability to automatically learn hierarchical representations has allowed CNNs to achieve strong performance in various computer vision tasks, including image classification and object recognition (Krizhevsky et al., 2017; He et al., 2016).

The development of more advanced CNN architectures has further improved their ability to extract detailed visual features. For example, AlexNet demonstrated the effectiveness of deep CNNs by significantly improving image classification performance through the use of multiple convolution layers and large-scale training data (Krizhevsky et al., 2017). Similarly, architectures such as ResNet introduced methods that allowed networks to become deeper while maintaining effective learning, improving the ability of models to capture complex visual patterns (He et al., 2016). These advancements have made CNNs increasingly useful for applications where visual information contains important predictive features.

In the context of real estate valuation, property images contain valuable information that may not be captured through traditional structured data alone. Features such as interior design, room organization, renovation quality, architectural style and overall property condition can influence how potential buyers perceive a home's value, yet these characteristics are difficult to quantify using traditional housing attributes, leaving a gap in conventional house price models. CNNs provide a method for extracting these visual characteristics and converting them into numerical representations that can be used for prediction models.

One of the earliest applications of CNNs in real estate prediction was explored by Poursaeed et al. (2017), who investigated vision-based real estate price estimation using property images. Their approach demonstrated that images could provide meaningful information for predicting property values, suggesting that visual features contain patterns related to housing prices. However, image-based approaches alone are limited because they do not include important measurable characteristics of a property, such as size, location and other structured attributes.

More recent research has expanded on image-based prediction by exploring how CNN-generated features can be combined with additional sources of information. Nguyen et al. (2023) examined the use of visual and textual features for real estate price prediction, demonstrating the potential benefits of combining different types of data. By using CNNs to process visual information while incorporating additional property characteristics, models can develop a more complete understanding of a property. This approach moves beyond simply recognizing visual patterns and instead focuses on creating more accurate property representations.

Although CNNs have demonstrated strong capabilities in extracting visual features, several challenges remain when applying them to real estate prediction. Unlike image classification tasks where the goal is to identify specific objects or categories, house price prediction requires understanding how visual characteristics relate to economic value. A model may successfully identify features of a home's appearance but still struggle to determine how

those features influence market price. Additionally, visual data alone cannot fully represent factors such as neighborhood quality, market conditions or geographic location.

Therefore, while CNNs provide a powerful method for extracting visual information from property images, accurate house price prediction also relies on structured housing data that capture measurable property characteristics. These structured features have long formed the foundation of traditional valuation models providing the quantitative information needed to estimate property values and serving as the basis for many modern prediction approaches (Geerts et al., 2023).

Structured Data in House Price Prediction

Before the introduction of deep learning and computer vision methods, house price prediction models primarily relied on structured data (Geerts et al., 2023). Structured data refers to measurable characteristics of a property that can be organized into numerical or categorical variables such as location, square footage, number of bedrooms, number of bathrooms, and other property attributes. These features have historically served as the foundation for automated valuation models because they provide consistent and quantifiable information that can be analyzed by predictive algorithms (Geerts et al., 2023).

Traditional approaches to real estate have shown that structured features can provide valuable insights into property prices (Geerts et al., 2023). Location, for example, is often considered one of the most influential factors because properties in different neighborhoods can have significantly different values despite having similar physical characteristics. Other variables, such as property size, age and number of rooms, also provide important information that helps models estimate housing prices. Because these features are directly connected to measurable aspects of a property, structured data has remained a reliable component of real estate prediction models (Geerts et al., 2023).

Machine learning methods have further improved the ability to analyze structured housing data by identifying complex relationships between property characteristics and market values (Kang et al., 2020). Researchers have explored various algorithms to predict house prices by learning patterns within large datasets containing historical sales information and property attributes (Kang et al., 2020). For example, studies using machine learning and geographic data have demonstrated that combining multiple structured sources can improve predictions by capturing relationships between properties and their surrounding environments (Kang et al., 2020). These approaches highlight the importance of structured information in understanding the economic factors that influence housing prices.

However, relying on structured data presents several limitations. Although numerical features provide important information, they often fail to capture qualitative characteristics that influence how buyers perceive a property. Factors such as interior design, architectural style, natural lighting, renovation quality, and overall appearance are difficult to represent through variables alone. Two homes with similar square footage and locations may have different values due to differences in condition or visual appeal, which structured data alone cannot fully capture.

This limitation has encouraged researchers to explore additional sources of information that can complement traditional housing features (Nguyen et al., 2023). While structured data provides measurable characteristics of a property, visual data offers insight into features that are more difficult to quantify. The combination of these two forms of information creates an opportunity for models to develop a more complete representation of a property by considering both numerical characteristics and visual appearance (Baltrušaitis et al., 2019).

As a result, recent research has shifted toward multimodal approaches that integrate structured data with visual information extracted through CNNs. By combining the strengths of both data types, these models aim to overcome the limitations of relying on either images or structured features alone and improve the accuracy of house price prediction.

Multimodal Learning for Visual and Structured Data Integration

While CNNs are effective at extracting meaningful features from images and structured data provides valuable information about measurable property characteristics, each data type has limitations when used independently. Visual data can capture details such as appearance, designs, and condition while structured data provides information such as location, size and other numerical characteristics. To overcome the limitations of using a single type of data, researchers have increasingly explored multimodal learning approaches that combine information from multiple sources (Baltrušaitis et al., 2019).

Multimodal learning focuses on developing models that can process and integrate different types of information to create a more complete representation of an object or environment (Baltrušaitis et al., 2019). In the context of house price prediction, multimodal models typically use CNNs to extract visual features from property images while simultaneously processing structured features through other machine learning methods (Nguyen et al., 2023). By combining these two forms of information, models can consider both the measurable characteristics of a property and the visual factors that influence its perceived value (Nguyen et al., 2023).

One common approach for multimodal house price prediction is feature fusion, where visual features extracted by CNNs are combined with structured property information before making a final prediction. In this approach, CNNs transform images into numerical feature representations that can be merged with variables such as location, square footage and property characteristics. This allows the model to learn relationships between visual patterns and housing attributes rather than relying on one source of information alone. Research in multimodal machine learning suggests that effective feature fusion can improve prediction performance by allowing models to capture complementary information across different data types (Baltrušaitis et al., 2019).

More recent research has explored combining visual information with additional property features to address this limitation. Nguyen et al. (2023) investigated real estate price prediction using both textual and visual features, demonstrating the benefits of integrating multiple sources of information. By combining features extracted from images with additional property

information, multimodal models can develop a more complete understanding of a property and potentially improve prediction accuracy compared to models using only one type of input.

Beyond traditional CNN-based methods, researchers have also explored newer architectures for multimodal real estate valuation. Yazdani and Raissi (2023) examined the use of self-supervised vision transformers for property valuation, demonstrating how advancements in computer vision continue to expand the ability of models to extract useful information from images. Although CNNs remain widely used, these developments show that the field is moving toward increasingly advanced methods for learning from complex visual data (Yazdani & Raissi, 2023).

Overall, research suggests that integrating visual and structured data through CNN-based multimodal models can improve house price prediction by capturing information that traditional approaches may overlook. However, the effectiveness of these approaches varies depending on the dataset, model architecture and fusion strategy used. Understanding these limitations is essential for developing more reliable and accurate real estate valuation systems.

Challenges and Future Directions

Although CNN-based multimodal approaches have demonstrated promising results in house price prediction, several challenges continue to limit their effectiveness and widespread adoption. Integrating visual and structured data is a complex process because the two types of information have different characteristics. Images contain large amounts of high-dimensional information, while structured data typically consists of a small number of measurable variables. Creating models that can effectively balance these different sources of information remains an important challenge in multimodal learning (Baltrušaitis et al., 2019).

Despite the improvements achieved through multimodal approaches, the effectiveness of these models depends heavily on how different types of data are combined. Simply adding image features and structured variables together does not guarantee better predictions. Researchers must determine which features provide meaningful information, how much influence each data type should have, and how differences between visual and numerical representations can be balanced. These challenges demonstrate that successful multimodal learning requires careful design of both the model architecture and feature integration methods.

One major challenge involves feature representation and fusion. CNNs are capable of extracting thousands of visual features from property images but not all extracted features contribute equally to predicting housing prices. Demonstrating which visual patterns are meaningful and how they relate to measurable property characteristics requires careful model design. Additionally, structured features such as location and square footage may have a stronger influence on housing prices than certain visual characteristics, creating the possibility that one type of data may dominate the other during training. Multimodal learning research has identified similar challenges involving feature alignment and fusion when combining different forms of information (Baltrušaitis et al., 2019).

Another challenge is the availability and quality of datasets. Deep learning models require large and diverse datasets to learn accurate patterns but real estate datasets can vary significantly in terms of location, property types, image quality and available features. A model trained on properties from one region may not perform as effectively when applied to another market with different housing trends and economic conditions. Additionally, inconsistencies in property images, missing information or limited historical data can affect model reliability.

Interpretability also remains an important concern for CNN-based house price prediction models. While these models can identify complex patterns and produce accurate predictions, understanding why a model reaches a specific valuation can be difficult. In real estate, where decisions involve significant financial investments, users may require explanation behind predictions rather than only receiving an estimated price. Future research may need to focus on developing more interpretable models that can highlight which visual and structured features contribute most to a prediction.

Despite these challenges, advancements in deep learning continue to create new opportunities for improving multimodal house price prediction. Researchers are exploring more advanced architectures including vision transformers and self-supervised learning methods, which may improve the ability of models to learn from limited or complex datasets (Yazdani & Raissi, 2023). Additionally improvements in data collection and feature integration methods may allow models to incorporate more diverse sources of information such as geographic data, market trends and additional property characteristics.

Overall, while CNN-based multimodal models have shown the potential to improve house price prediction by combining visual and structured information, challenges related to data quality, feature integration and model interpretation remain. Addressing these limitations will be essential for developing more accurate, reliable and practical real estate valuation systems in the future.

Conclusion

Convolutional neural networks have transformed the way researchers approach visual data by allowing models to automatically identify complex patterns and extract meaningful features from images. In the context of house price prediction, CNNs provide valuable information by capturing visual characteristics of properties that traditional structured data alone cannot represent such as architectural style, property conditions and overall appearance. However, research suggests that CNNs are most effective when combined with structured housing information, including location, square footage and other measurable property features. By integrating these two sources of information, multimodal models can create a more complete representation of a property and improve upon the limitation of using either visual data or structured data independently.

Works Cited

Ahmed, Eman, and Mohamed N. Moustafa. “House Price Estimation from Visual and Textual Features.” arXiv, 2016, arxiv.org/abs/1609.08399.

Baltrušaitis, Tadas, et al. “Multimodal Machine Learning: A Survey and Taxonomy.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, 2019, pp. 423-43. https://doi.org/10.1109/TPAMI.2018.2798607.

Geerts, Mitchell, et al. “A Survey of Methods and Input Data Types for House Price Prediction.” ISPRS International Journal of Geo-Information, vol. 12, no. 5, 2023, article 200. https://doi.org/10.3390/ijgi12050200.

He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-78. https://doi.org/10.1109/CVPR.2016.90.

Kang, Jinmyung, et al. “Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning.” Landscape and Urban Planning, vol. 202, 2020, article 103730. https://doi.org/10.1016/j.landurbplan.2019.103730.

Krizhevsky, Alex, et al. “ImageNet Classification with Deep Convolutional Neural Networks.” Communications of the ACM, vol. 60, no. 6, 2017, pp. 84-90. https://doi.org/10.1145/3065386.

LeCun, Yann, et al. “Deep Learning.” Nature, vol. 521, no. 7553, 2015, pp. 436-44. https://doi.org/10.1038/nature14539.

Nguyen, Trong V., et al. “Real Estate Pricing Prediction via Textual and Visual Features.” Machine Vision and Applications, vol. 34, 2023, article 74. https://doi.org/10.1007/s00138-023-01464-5.

Poursaeed, Omid, et al. “Vision-Based Real Estate Price Estimation.” arXiv, 2017, arxiv.org/abs/1707.05489.

Saxena, Ansh. “An Introduction to Convolutional Neural Networks.” International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 12, 2022, pp. 943-47. https://doi.org/10.22214/ijraset.2022.47789.

Simonyan, Karen, and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” International Conference on Learning Representations, 2015, arxiv.org/abs/1409.1556.

Srivastava, Nitish, and Ruslan Salakhutdinov. “Multimodal Learning with Deep Boltzmann Machines.” Journal of Machine Learning Research, vol. 15, 2014, pp. 2949-80. jmlr.org/papers/v15/srivastava14a.html.

Szegedy, Christian, et al. “Going Deeper with Convolutions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594.

Wu, Yuan, and Yan Zhang. “House Price Prediction Based on Deep Learning.” arXiv, 2022, arxiv.org/abs/2204.09050.

Yazdani, Mohsen, and Maziar Raissi. “Real Estate Property Valuation Using Self-Supervised Vision Transformers.” arXiv, 2023, arxiv.org/abs/2302.00117.

Potential of Cell Towers By Vaibhav Gaddam

Abstract

This paper covers current issues with meteorological detection and observation of tornadoes. It provides potential ways to address such issues through the use of cell towers. This paper discusses the physics of how radio signals are affected by tornadoes and how this can help overcome current issues with Doppler radar for imaging and recording data from tornadoes. Another point this paper covers is how the use of cell towers for monitoring tornadoes can help improve early tornado detection speed and provide better clarity of tornado evacuations.

Keywords: Tornado, Doppler Radar, Cell Tower, Phase Shift, Degradation, Mie Scattering, Microwaves

Introduction

Current methods of monitoring tornadoes involve the use of Doppler radars. These help create models and images of tornadoes for further study, but are lacking in certain aspects. Evidence of this can be seen in the fact that Doppler radars have a hard time getting information on low altitude regions of tornadoes but also have patchy data on the inside of tornadoes. Additionally, the use of Doppler radar takes about 3 to 5 minutes to fully scan the different atmospheric levels of a tornado along with needing a meteorologist to predict where a tornado will be based on its rotation if the tornado's visuals are lost. One possible solution to this issue is to use cell towers. Cell towers have already been used to help determine when tornadoes could be formed, but they can also assist with getting further information on tornadoes, especially their dynamics at low altitudes and the inside of them. Moreover, the use of cell towers improves the early detection of tornadoes, as once the cell towers in a network all experience similar attenuations in their signal, meteorologists will know where a tornado is, allowing for improvement in evacuation notice.

LITERATURE REVIEW

1. Current Meteorological Detection

Doppler radar systems like NEXRAD used by the NOAA make up the core of weather detection using algorithms like Tornadic Vortex Signature to identify atmospheric rotations like tornadoes. Doppler radars climb in altitude as they leave a weather station, so they take measurements of altitudes of several kilometers above ground ("Severe Weather 101").

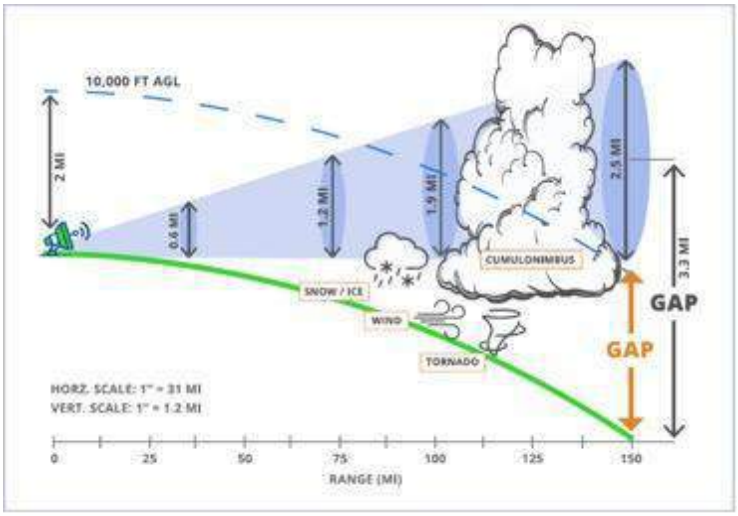


Figure 1. Radar beam coverage gap due to Earth's curvature. National Weather Service, National Oceanic and Atmospheric Administration.

This makes it so that trying to gather data on the lower altitude regions of a tornado is difficult and lacks clarity and resolution.

To bypass the altitude limit from Doppler radars, meteorologists have tried implementing mobile weather balloons to scan the internal funnel of a tornado. The core of a tornado has a large congregation of water droplets and hail. Microwaves are roughly the same size as the raindrops and hail, so the microwaves undergo Mie scattering, leading to a massive drop in the power of the waves and leading to less clarity in imaging the center of the tornado (Darling).

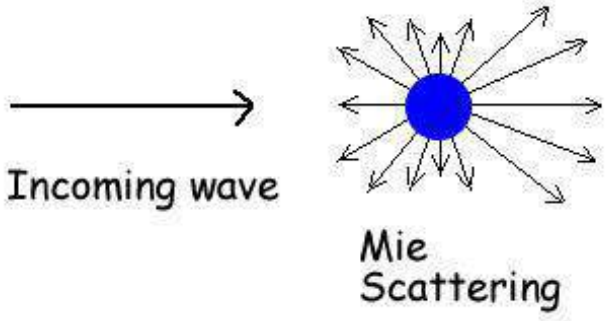


Figure 2. Diagram illustrating Mie scattering. Adapted from "Mie Scattering," by D. Darling, 2016, The Internet Encyclopedia of Science.

The signal refracts and scatters into multiple directions, blurring the images and data, making it extremely difficult to analyze information about a tornado's internal structure.

2. Mechanism of Cell Tower Detection

To maintain signal connections, cell towers keep track of the phase shift of the signal and the overall signal strength ("Mobile Fronthaul"). Using this information, meteorologists can accumulate information about the storm and even determine if one is about to form. By

measuring this information, it is possible to determine things like the humidity of an area, which causes a greater phase shift and signal degradation (Zhang et al.). An extension of this idea is that, instead of using it to identify just humidity, meteorologists may be able to identify the magnitude and vorticity of a tornado. This enables meteorologists to get more data on tornadoes and how their dynamics work at low altitudes and potentially find relations between changes in a tornado's vorticity, since a tornado's vorticity is not consistent throughout the entire storm.

Tornadoes also have irregular objects and asymmetric shapes, so waves do not travel or refract cleanly through them, and when a microwave from a cell tower hits the debris cloud, the wave scatters. The wave splits into many fragmented paths that bounce off the debris due to a mix of refraction, reflection and diffraction, causing the signal to arrive at the cell tower with different phase shifts and at slightly different time intervals (Poole). The cell tower will experience rapid fluctuations in power called small-scale fading, which is essentially radio static (Durgin and Rappaport). The amount of fading that occurs can correlate to the density and size of the debris cloud, giving a rough estimate of the tornado's magnitude. Moreover, the inside of a tornado is an extremely low pressure zone and acts like a vacuum. Due to the decrease of density in the eye of the tornado, its refractive index is smaller than the refractive index of the funnel, so any radio waves that go through the entirety of a tornado will slow down, speed up, and then slow back down all while bending through the change in mediums of the funnel and the eye.

This rapid fluctuation in the signal speed and angle will cause a significant alteration in the phase shift of the wave along with the intensity. When cell towers record this phase shift, meteorologists can use this information to try and approximate the pressure drop inside the tornado based on the shift recorded, along with other information like the polarization and intensity of the received wave compared to the initial signal sent. Unlike in radar, this is not detrimental for cell towers since the signal is being measured after it goes through the tornado rather than measuring a signal being reflected like radar. The degradations of the signal are what could enable meteorologists to further identify properties of a tornado.

3. Combination of Cell Towers and Current Meteorological Detection

The use of cell towers in conjunction with current meteorological imaging like Doppler radar can provide more information and insight on tornadoes. The NEXRAD Doppler radar networks used by the NOAA provide high-altitude data of the tornado, while the use of cell towers makes up for that gap as they can provide valuable data on low altitude interaction of a tornado.

Additionally, the use of cell towers may be able to collect more information on the inside of the tornado. This is because with cell towers, the signal is being measured after it goes through the tornado, while Doppler radar measures a reflected signal, so it cannot gather as much information on the inside of a tornado. This helps meteorologists overcome the issues they face due to time scattering, which causes the reflected signal to be out of phase and degraded, while for cell towers this is beneficial. Meteorologists can use the information given by cell towers in coordination with the current data they already have to make more accurate and sophisticated

models of tornadoes. Cell towers are already readily available throughout the U.S.A., so leveraging the current grid of cell towers can help improve tornado detection and understanding. Furthermore, cell towers could be used like a network of trip wires for a tornado, since, if the signals between cell towers start to experience a similar kind of degradation, it can help notify meteorologists about the location of a tornado, and could help keep people more informed about a tornado's path if it ever deviates from its predicted path. This allows pedestrians in a tornado's line of sight to have not only early evacuation notice but also more accurate evacuation notice since, through this trip-wire-like monitoring, meteorologists will also have a relative size and magnitude of a tornado so there is less guesswork in what areas a tornado might impact. Meteorologists have a hard time finding this information since they cannot access much information about tornadoes from 100 meters to ground level while cell towers can.

4. **Conclusion**

The integration of cell towers as a way to supplement current meteorological detection like Doppler radars can greatly enhance how meteorologists can study tornadoes. It can help provide more data on tornadoes at low altitudes and even give more insight into how the core of a tornado behaves. When paired with current radar technology, meteorologists will have an easier time collecting data and modeling tornadoes. This could prove a great benefit in improving the current understanding of how Earth's atmospheric systems work and provide a way to improve tornado evacuation by allowing residents more notice and accuracy for an evacuation, leading to more lives saved from a tornado.

Works Cited

Darling, David. "Mie Scattering." The Internet Encyclopedia of Science, 2016, www.daviddarling.info/encyclopedia/M/Mie_scattering.html.

Durgin, Gregory D., and Theodore S. Rappaport. "Theory of Multipath Shape Factors for Small-Scale Fading Wireless Channels." IEEE Transactions on Antennas and Propagation, vol. 48, no. 5, May 2000, pp. 682–93. Georgia Institute of Technology, propagation.ece.gatech.edu/ECE3065/tutorials/Project04/project/TransAP00.pdf.

"Mobile Fronthaul: Remote Radio Head Systems." EITC Technology Roadmap, EITC, eitc.org/research-opportunities/5g-and-beyond-mobile-wireless-technology/5g-and-beyond-technology-roadmap/the-fusion-of-fronthaul-and-backhaul/mobile-fronthaul/remote-radio-head-systems/remote-radio-head-systems.

National Weather Service. Radar beam coverage gap due to Earth's curvature. National Oceanic and Atmospheric Administration, NWS Radar.

"Severe Weather 101: Tornado Detection." National Severe Storms Laboratory, National Oceanic and Atmospheric Administration, www.nssl.noaa.gov/education/svrwx101/tornadoes/detection/.

"Study Turns Cell Phone Towers Into Rain Gauges." Climate Central, 4 Feb. 2013, www.climatecentral.org/news/study-turns-cell-phone-towers-into-rain-gauges-15552.

Zhang, Pingping, et al. "A Review on Rainfall Measurement Based on Commercial Microwave Links." Sensors, vol. 22, no. 12, 2022, p. 4395. MDPI, www.mdpi.com/1424-8220/22/12/4395.

Poole, Ian. "Multipath Radio Propagation: Wireless Signal Multipath." Electronics Notes, www.electronics-notes.com/articles/antennas-propagation/propagation-overview/multipath-propagation.php.

A Mathematical Proof of the Necessity of Individualized Dosing Intervals in Patients with Reduced Metabolism By JuneYoung Park

Abstract

This study used an exponential decay model and the formula for the sum of a geometric series to generalize the concentration of drugs in the body following repeated dosage and to mathematically determine optimal, patient-specific dosing intervals. Applying this model to cases of penicillin administration in healthy individuals and patients with chronic kidney disease (CKD) demonstrated that CKD patients with decreased metabolic function require dosing intervals of approximately 10 to 40 hours—significantly longer than the standard 3-hour interval—to prevent drug accumulation and toxic side effects. The study is significant for providing a clinical basis by mathematically proving the necessity of designing individualized drug dosage regimens for patients with reduced metabolism.

Keywords: Drug Concentration, Exponential Decay, Geometric Series, Dosing Interval, Reduced Metabolism

1. Introduction

In modern medicine, maximizing therapeutic efficacy while minimizing side effects is a key challenge in safe drug dosage. For a drug to be effective within the body, its concentration must exceed the effective level but remain below the toxic level. Establishing a precise dosing time interval is essential to ensure that drug levels remain within the range during continuous dosage. If the interval is too long, the concentration in the body will drop under the effective level before the next dosage. If the interval is too short, the concentration will rise to the toxic level and cause side effects.

Therefore, this paper aims to derive a mathematical model that generalizes the maximum and minimum drug concentrations in the body during repeated dosage by utilizing an exponential decay model and the formula for the sum of a geometric series. Furthermore, by applying this model to cases of penicillin dosage in healthy individuals versus patients with chronic kidney disease, the study seeks to mathematically demonstrate why establishing individualized dosing intervals is essential for patients with reduced metabolic function and to discuss the associated medical implications.

2. Background

2.1 Basic Pharmacokinetic Model

According to Rowland and Tozer (31-45), the basic pharmacokinetic model explaining the decay of drug concentration in the body is an exponential function:

$$C(t) = C_0 \cdot e^{-kt}$$

where t is the elapsed time (in hours, h), and C(t) and C_0 represent the drug concentration in the

body at time t and the initial concentration immediately after administration, respectively (in milligrams per liter, mg/L), as shown in Fig. 1. In this paper, however, we replace e^k with decay factor a, where a > 1, redefining the model as:

$$C(t) = C_0 \cdot a^{-t}$$

This substitution is utilized for the ease of calculation.

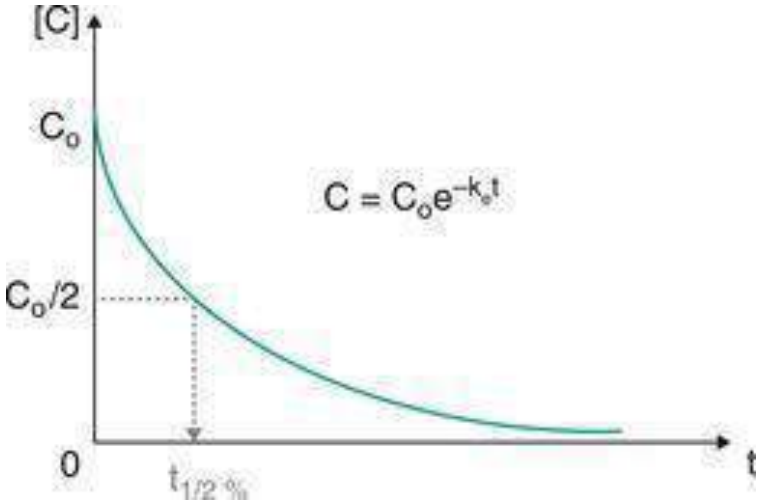


Fig. 1. A graph showing the exponential decay of drug concentration in the body after a single dose. ("6: Les caractéristiques...")

2.2 Therapeutic Window

A drug is effective when its concentration is over MEC (Minimum Effective Concentration) and is safe when its concentration is below MTC (Maximum Toxic Concentration). This range is called a therapeutic window—the range of drug concentration that can treat a disease effectively without causing unsafe toxic side effects (Brunton and Knollmann 12-15). A proper dosing time interval should maintain the concentration over MEC and below MTC.

2.3 Geometric Series

A geometric series is the sum of the terms of a geometric sequence, where each successive term is obtained by multiplying the preceding term by a constant called the common ratio (r). The general mathematical representation of a term of geometric sequence is given by:

$$a_n = ar^{n-1}$$

The general mathematical representation of a geometric series is therefore:

$$S_n = a + ar + ar^2 + \ldots + ar^{n-1}$$

The formula for the sum of geometric series is:

$$S_n = \frac{a(1-r^n)}{1-r}$$

3. Body

3.1 Finding the Range of Proper Dosing Interval

As the drug is dosed at certain time intervals, its concentration in vivo repeats increasing and decreasing. Generally, the concentration of the drug in vivo is the highest just after the dose and is the lowest just before the next one.

First, the maximum concentration of the drug in vivo can be determined as follows. Let the dosing time interval be T hours. The concentration of the drug in vivo just after the first dose is equal to the initial concentration, which is:

$$C_0 \cdot a^{-0} = C_0$$

If the drug is dosed for the second time after T hours, the concentration is equal to the initial concentration added by the remaining concentration, which is:

$$C_0 + C_0 \cdot a^{-T} = C_0\left(1 + a^{-T}\right)$$

Likewise, the concentration after the third dose is:

$$C_0 + \left(C_0 + C_0 \cdot a^{-T}\right) \cdot a^{-T} = C_0\left(1 + a^{-T} + a^{-2T}\right)$$

Therefore, the concentration after n^{th} dose can be generalized by the following expression:

$$C_0\left(1 + a^{-T} + a^{-2T} + \ldots + a^{-(n-1)T}\right)$$

Let this value be defined as C_{max}. Using the formula for the sum of geometric series, the expression can be simplified as:

$$C_{max} = C_0(1 + a^{-T} + a^{-2T} + \ldots + a^{-(n-1)T}) = \frac{C_0\left(1-a^{-nT}\right)}{1-a^{-T}}$$

This shows that the value of C_{max} is not fixed and changes by the value of n. However, C_{max} converges to a certain value as n keeps increasing. Since the drug is repeatedly dosed, it can be assumed that C_{max} eventually reaches the convergence value. This state is called SS (Steady State), and let the maximum concentration in Steady State be defined as $C_{max,ss}$ (Shargel and Yu 180-195). The expression for $C_{max,ss}$ can be found by taking the limit of C_{max}:

$$C_{max,ss} = lim_{n\to\infty} C_{max} = lim_{n\to\infty} \frac{C_0\left(1-a^{-nT}\right)}{1-a^{-T}} = \frac{C_0}{1-a^{-T}}$$

The minimum concentration of the drug in vivo can be found in the same way. The concentration of the drug just before the second dose is:

$$C_0 \cdot a^{-T}$$

The concentration just before the third dose is:

$$\left(C_0 + C_0 \cdot a^{-T}\right) \cdot a^{-T} = C_0 \cdot (a^{-T} + a^{-2T})$$

Likewise, the concentration just before the n^{th} dose can be generalized by the following expression:

$$C_0 \cdot (a^{-T} + a^{-2T} + \ldots + a^{-(n-1)T})$$

Let this value be defined as C_{min}. The expression can be simplified using the formula for the sum of geometric sequence:

$$C_{min} = C_0\left(a^{-T} + a^{-2T} + \ldots + a^{-(n-1)T}\right) = \frac{C_0 \cdot a^{-T}\left(1-a^{-(n-1)T}\right)}{1-a^{-T}}$$

Like the maximum concentration C_{max}, C_{min} converges to a certain value as n keeps increasing (Shargel and Yu 180-195). Let the minimum concentration in steady state be defined as $C_{min,ss}$:

$$C_{min,ss} = lim_{n\to\infty} C_{min} = lim_{n\to\infty} \frac{C_0 \cdot a^{-T}\left(1-a^{-(n-1)T}\right)}{1-a^{-T}} = \frac{C_0 \cdot a^{-T}}{1-a^{-T}}$$

In order to meet the therapeutic window, $C_{max,ss}$, $C_{min,ss}$ should satisfy the following inequalities:

$$C_{max,ss} < MTC,\ C_{min,ss} > MEC$$

$$\frac{C_0}{1-a^{-T}} < MTC,\ \frac{C_0 \cdot a^{-T}}{1-a^{-T}} > MEC$$

The proper dosing interval T should therefore be:

$$log_a\left(\frac{MTC}{MTC-C_0}\right) < T < log_a\left(\frac{C_0+MEC}{MEC}\right)$$

3.2 Applying the Model to Patients with Reduced Metabolism: The Case of Penicillin

In case of penicillin, when dosed to healthy individuals, the decay factor has a value of approximately $a_{healthy} = 1.8$, calculated based on pharmacokinetic parameters (Katzung and Vanderah, ch. 43). In this paper, we take 3g of penicillin as an example, of which the initial concentration in vivo (C_0) is 150mg/L. Due to the wide therapeutic window of penicillin, it can be hypothetically assumed that the MEC and MTC of penicillin are 25mg/L and 400mg/L, respectively. The range of proper dosing intervals is therefore:

$$0.799 < T_{healthy} < 3.310$$

In the pharmaceutical industry, the dosing interval is usually chosen as an integer. It can be assumed that the dosing interval of penicillin is $T_{healthy} = 3$.

Now this result ($T_{healthy} = 3$) can be applied to patients with CKD (chronic kidney disease), which is a well-known metabolic disorder. According to Munar and Singh, CKD reduces kidney function and consequently delays the elimination of the drug (1487). Thus, the decay factor of patients with CKD (a_{CKD}) is smaller than $a_{healthy} = 1.8$.

$$1 < a_{CKD} < a_{healthy}$$

Therefore, in this paper, we assume that a_{CKD} is 1.05. In case of CKD patients, the value of $C_{max,ss}$ and $C_{min,ss}$ are each:

$$C_{max,ss} = \frac{C_0}{1-a^{-T}} = 1101.626$$

$$C_{min,ss} = \frac{C_0 \cdot a^{-T}}{1-a^{-T}} = 951.626$$

Since the MEC of penicillin is 25, the value of $C_{min,ss}$ satisfies the condition. However, the value of $C_{max,ss}$ exceeds the MTC, which does not meet the therapeutic window. Indeed, the range of individualized dosing interval of penicillin for patients with CKD is:

$$9.633 < T_{CKD} < 39.883$$

Therefore, patients with CKD should take penicillin at much longer intervals, specifically between approximately 10 to 40 hours, rather than the standard 3-hour interval. This proves that the proper dosing interval for healthy individuals may not fit the ones with reduced metabolism.

4. Conclusion

In this paper, we calculated the range of proper dosing intervals using an exponential decay model and applied it to healthy people and patients with chronic kidney disease. By analyzing the drug concentration at Steady State through geometric series, the optimal dosing interval for healthy individuals was determined to be approximately 3 hours ($T_{healthy} = 3$). In contrast, patients with severe CKD, whose drug elimination is significantly delayed due to impaired metabolic function ($a_{CKD} = 1.05$) required a drastically extended interval of approximately 10 to 40 hours. This clear discrepancy mathematically demonstrates that standard dosing regimens may be unsuitable for patients with reduced metabolism and underscores the necessity of designing individualized, patient-specific drug dosage regimens in clinical settings.

If a dosing interval does not satisfy the time range, which we calculated, the concentration in vivo will fall outside the therapeutic window. If the concentration drops below the lower limit, the MEC, the drug does not have enough therapeutic effect. If the concentration exceeds the upper limit, the MTC, the drug can have toxic side effects that can harm the drug recipient. In case of penicillin, when the concentration in vivo is over its MTC, which is 400mg/L, it can penetrate the BBB (Blood-Brain Barrier) and enter the brain cells. As a result, it can cause neurotoxic side effects such as seizures, delirium, or myoclonus, and can end up in a coma. Given the potential risks of overdose not only with penicillin but also with many other drugs, precise and individualized dosing interval calculation for patients with reduced metabolism is essential.

Works Cited

Brunton, Laurence L., and Björn C. Knollmann, editors. *Goodman & Gilman's: The* Pharmacological Basis of Therapeutics. 14th ed., McGraw Hill, 2023.

Katzung, Bertram G., and Todd W. Vanderah, editors. Basic & Clinical Pharmacology. 15th ed., McGraw Hill, 2021.

Munar, M. Y., and Harleen Singh. "Drug Dosing Adjustments in Patients with Chronic Kidney Disease." American Family Physician, vol. 75, no. 10, 15 May 2007, pp. 1487–96, www.aafp.org/afp/2007/0515/p1487.

"6: Les caractéristiques pharmacocinétiques des médicaments après administration par voies intravasculaires et extravasculaires." Medicine Key, 4 May 2017, clemedicine.com/6-les-caracteristiques-pharmacocinetiques-des-medicaments-apres-administration-par-voies-intravasculaires-et-extravasculaires/. Accessed 25 July 2026.

Rowland, Malcolm, and Thomas N. Tozer. Rowland and Tozer's Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications. 5th ed., Wolters Kluwer, 2019.

Shargel, Leon, and Andrew B. C. Yu. Applied Biopharmaceutics & Pharmacokinetics. 8th ed., McGraw Hill, 2022.

The Effect of Graphene Oxide Concentration on the Growth Dynamics and Apparent Amoxicillin Susceptibility of *Lactobacillus acidophilus* During Serial Subculture

By Tengwen Lou and Xinyi Liu

Abstract

This experiment investigated whether the sensitivity of a yogurt probiotic bacteria found in the gut, *L. acidophilus, to the* antibiotic amoxicillin is influenced under the conditions of graphene oxide. Two batches of five samples each with bacteria and GO are used, one with antibiotic exposure and one without. The results showed that GO combined with amoxicillin has a significant impact on growth dynamics and apparent amoxicillin susceptibility, which inhibits its growth until five tested generations and down to 10.1%. High GO concentrations appear to reduce growth inhibition in antibiotic-free conditions —potentially due to nanomaterial aggregation—whereas combined exposure with amoxicillin markedly suppresses bacterial coverage, which rebounded to 88.7%. Given the expanding use of graphene oxide in nanotechnology as well as biomedical and environmental fields, this experiment helps to understand its relation with a common type of probiotic bacteria, which is important for assessing its potential impacts relating to microbiome stability.

Introduction

Antibiotic resistance is a major challenge in microbiology because bacteria can acquire heritable changes that reduce their susceptibility to antimicrobial drugs. *Lactobacillus acidophilus* is a Gram-positive lactic acid bacterium with an optimal growth temperature of 37 °C, commonly associated with fermented dairy products and often found in gut lining, aiding digestive health and grows well in MRS medium. (Gao, 2022) Meanwhile, Graphene oxide (GO) is a two-dimensional carbon-based nanomaterial with oxygen-containing functional groups that give it distinct physicochemical properties. It is used in functions such as drug delivery and water purification. (Anegbe, 2024) However, reduced bacterial growth inhibition following antimicrobial exposure does not necessarily demonstrate antibiotic resistance. Short-term antimicrobial effects may instead produce physiological stress or cell damage, while phenotypic tolerance can allow bacteria to survive exposure without acquiring a heritable increase in resistance.

This experiment investigates the question of "How does graphene oxide concentration (0μg/mL, 10μg/mL, 20μg/mL, 40μg/mL, 80μg/mL) affect the growth response of *L. acidophilus* to amoxicillin following serial subculture, as quantified by colony-area coverage (%) using image analysis?" Two batches with each five samples of bacteria with antibiotics were used, one with GO exposure and one without, and bacterial colonies were observed and measured to compare growth.

The toxicity of graphene oxide (GO) to Gram-positive bacteria mainly stems from physical cleavage at its sheet edges, contact-mediated membrane damage, and induced oxidative stress. (Akhavan, 2010) However, most relevant data come from pathogenic bacteria such as

Staphylococcus aureus, while research on beneficial bacteria such as Lactobacillus acidophilus remains limited. Amoxicillin weakens bacterial cell walls by inhibiting peptidoglycan cross-linking, exhibiting bactericidal activity against Gram-positive bacteria in the division phase. Under GO co-existence conditions, impaired cell wall integrity may further increase the sensitivity of bacteria to physical damage from nanomaterials. (Akhavan, 2010)

Harmless bacteria were chosen instead of the infectious bacteria because it provides us with a new perspective. Understanding how GO affects beneficial bacteria could provide insight into its impact on the gut microbiome and probiotic effectiveness if it comes into contact when antibiotics are present. Both GO and amoxicillin have antimicrobial properties, which provides the possibility that their combined presence could either enhance, reduce, or even alter bacterial susceptibility. (Liu, 2011)

Given the expanding use of graphene oxide in nanotechnology as well as biomedical and environmental fields, this experiment helps to understand its relation with a common type of probiotic bacteria, which is important for assessing its potential impacts relating to microbiome stability. If this is proven effective, that GO does alter antibiotic susceptibility in beneficial bacteria, it can introduce a whole new layer of insight on the gut microbiome, probiotic effectiveness when mixed with antibiotic, and the influence of nanomaterials on health.

The null hypothesis will be that the GO concentration will not significantly affect colony-area coverage, regardless of amoxicillin exposure. Alternatively, the hypothesis will be increasing GO concentration (0–80 μg/mL) will alter *L. acidophilus* colony-area coverage, with prediction that higher GO concentrations will generally reduce colony-area coverage in the amoxicillin-containing group due to the combined antimicrobial effects. Such a hypothesis is made because GO can inhibit bacterial growth through surface interactions, membrane disruption, and oxidative stress (Liu, 2011). Amoxicillin also inhibits bacterial cell-wall synthesis (Charteris, 1998). Lactobacillus acidophilus is generally susceptible to amoxicillin, although susceptibility is typically assessed by minimum inhibitory concentration (MIC) testing rather than colony-level growth measurements (Charteris, 1998). Therefore, their combined exposure is expected to produce greater growth inhibition. However, GO aggregation may alter its effective antibacterial activity, so the response may not be strictly linear.

Variables

Independent variable

Graphene oxide concentration of five levels: 0, 10, 20, 40, and 80 μg/mL. This variable was used to investigate the effect of different GO concentrations on bacterial adaptive responses.

Dependent variables

1. Colony coverage percentage
2. The proportion of colony area to the total plate area was quantitatively measured using an image analysis tool (automeris.io) to reflect bacterial growth density.

3. Colony color and morphology

Controlled variables

1. Bacterial strain and initial inoculum size.

All samples used the same batch of *Lactobacillus acidophilus* capsules, with the initial inoculum size controlled by the same mass dissolved in MRS broth to create bacterial suspension.

2. Culture medium type and volume.

MRS agar was used uniformly, and MRS broth all had a final volume of 20 mL per tube.

3. Antibiotic concentration.

The final concentration of amoxicillin in the antibiotic-containing group was fixed at 20 μg/mL.

4. Culture temperature and time.

All cultures were conducted in the same room temperature, with each generation lasting 24 hours.

5. Subculturing cycle and transfer volume.

1 mL of bacterial suspension was transferred to a fresh culture medium of the same concentration every 24 hours, for a total of 5 generations.

6. Culture environment (aerobic/anaerobic).

Strictly anaerobic equipment was not used in the experiment; all samples were treated under the same atmospheric conditions.

7. Spreading and image acquisition methods.

The same size petri dishes were used for coating, and images were acquired under the same light source and distance.

Methods

Part A – Preparation of Base Solutions

Freshly cultured *Lactobacillus acidophilus* was prepared. Its source and train designation is Lactobacillus acidophilus LA02, with the capsule composition consisting of cellulose, magnesium stearate and a vegetable cellulose capsule shell. Each capsule provides a viable cell count of 3 billion, at the time of manufacture.

MRS broth was prepared by mixing 54g powder with 1000 mL distilled water and sterilized according to lab protocols at 121°C. Agar jelly was created using 20g agar power, strength 1400g/cm², in 1000 mL distilled water, then 15 mL were placed into a 90 mm diameter petri dish.

A stock solution of graphene oxide (GO) was prepared at 400 μg/mL by diluting 32 mL of 1250 μg/mL GO concentrate to a final volume of 100 mL with sterile distilled water then thoroughly mixed to ensure even dispersion.

Amoxicillin stock solution was prepared to a concentration of 400 μg/mL and refrigerated until use.

Part B – Preparation of Test Tubes

A total of ten test tubes have been prepared, five containing amoxicillin of 20 μg/mL combined with rising concentrations of graphene oxide, and another five each with only rising concentrations of graphene oxide; all with a total volume of 20 mL solution. To maintain complete clarity regarding chemical dosages, detailed solution preparation protocols were prepared in Table 1

To prepare the amoxicillin group

1 mL of amoxicillin stock solution (400 μg/mL) was added to each of the five test tubes, then increasing concentrations of graphene oxide were added to reach the desired final concentration:

0 μg/mL 0 mL
10 μg/mL 0.5 mL
20 μg/mL 1.0 mL
40 μg/mL 2.0 mL
80 μg/mL 4.0 mL

Then, MRS broth (prepared in later steps) was added to reach a total volume of 20 mL. These procedures were repeated for 3 times to ensure technical repeat for 3 times.

To prepare the amoxicillin-free group

Instead of amoxicillin, distilled water of equal volume was added to the remaining five test tubes, followed by the same concentrations of graphene oxide as the amoxicillin group, from 0 μg/mL to 80 μg/mL.

Then, MRS broth (prepared in later steps) was added to reach a total volume of 20 mL.

Part C – Inoculation and Incubation

Freshly cultured *Lactobacillus acidophilus* was prepared. Its source and train designation is Lactobacillus acidophilus LA02, With the capsule composition consisting of cellulose, magnesium stearate and a vegetable cellulose capsule shell. Each capsule provides a viable cell count of 3 billion organisms at the time of manufacture.

A capsule of 500 mg *L. acidophilus* was added to 20 mL of distilled water, then mixed thoroughly to ensure equal dispersion.

Then, equal volume of the MRS broth was added to each of the ten test tubes to reach a total volume of 20 mL, then gently mixed and left to settle.

Afterwards, all tubes were incubated for 24 hours.

Part D – Subculturing (Development of Resistance)

After 24 hours of incubation, 1 mL of bacterial culture was removed from each tube and transferred to freshly prepared tubes containing the same concentrations of amoxicillin and higher concentration of GO.

This process was repeated for every 24 hours, each round equating to one generation.

5 generations were subcultured to observe potential resistance development.

Part E – Observational Stage

1 mL of bacterial culture was removed from each tube, including the replicated group, then spread onto an MRS agar plate, each repeat group occupying one-third of the area.

The plates were then incubated for 24 hours, then the percentage of colony area was calculated using the website tool automeris.io

Part F – Safety, Ethical & Environmental Considerations

Graphene oxide is non-hazardous unless ingested. *Lactobacillus acidophilus* is a strain of harmless gut bacteria. GO and amoxicillin were handled in rooms with good ventilation. Gloves, face masks, and safety goggles were worn throughout the experiment. All contaminated materials were disposed of following institutional protocols. No animal or human subjects were involved. Glassware was reused where possible to reduce waste.

Qualitative data

Figure 1: Growth results of the bacterial colony after five days
Top row: GO + Amoxicillin // Bottom row: Only GO
Each group consists of one top and one bottom dish counting left to right, 5 groups in total

1. Amoxicillin-Free Group

At 10–40 μg/mL GO, colony density slightly decreased, indicating mild inhibitory effects. At 80 μg/mL GO, colony density rebounded, suggesting bacterial adaptation under higher GO loads.

2. Amoxicillin-Containing Group

Colony density was significantly reduced across all groups, confirming strong antibiotic inhibition. At 10–40 μg/mL GO + amoxicillin, colonies showed pale red to reddish-brown discoloration caused by oxidative stress, indicating sublethal stress. At 40–80 μg/mL GO, colony color returned to normal milky white, reflecting reduced stress due to GO aggregation or adaptive responses.

3. Comprehensive Interpretation

These results reveal a non-monotonic adaptive response in *L. acidophilus*: lower GO concentrations increase sublethal stress (shown in sparse colonies, discoloration), while higher GO concentrations alleviate stress (shown in increased density, normal color). This suggests physiological adaptation, perhaps through gene modification/evolution over generations, rather than direct resistance, with GO effects dependent on its dispersion state and effective concentration.

Quantitative data

Table 1: Percentage of colony area (%) in each generation (shown in the image)

	Amoxicillin-Free Group (with GO)	Amoxicillin-Containing Group (with GO)
0μg/mL	92.7%	89.7%
10μg/mL	87.8%	81.6%
20μg/mL	82.8%	52.4%
40μg/mL	70.1%	49.6%
80μg/mL	88.7%	10.1%

Processed data

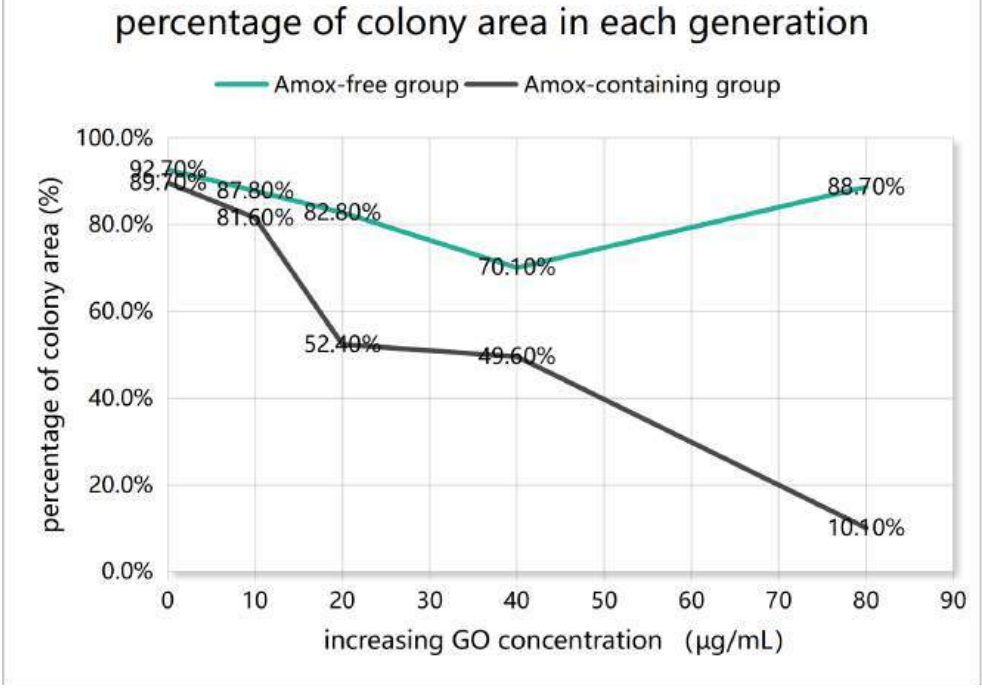


Figure 2: Graph of percentage of colony area in each group

Table 2: Descriptive statistics of data (%) under two conditions

	Amoxicillin-Free Group	Amoxicillin-Containing Group
Maximum value	92.7%	89.7%
Minimum value	70.1%	10.1%
Range of values	22.6%	79.6%

Table 3: Difference (%) and percentage difference (%) between each condition

Absolute percentage-point changes (Δ%) and relative percentage changes (%Δ) between consecutive GO concentration increments were calculated using standard mathematical formulations:

Percentage-point Difference (Δ%) $= Y_{i+1} - Y_i$

(Where i = the index of the current GO concentration level)

Worked example: concentration group 0 vs 10 μg/mL

Percentage-point Difference % = 87.8%−92.7% = -4.9%

eqX1: Equation for calculating Percentage-point Difference

Relative Percentage change (%Δ) = $(\frac{Y_{i+1}-Y_i}{Y_i}) \times 100\%$

Relative Percentage Change= (87.8-92.7)/92.7×100% = -5.3%

eqX2: Equation for calculating Relative Percentage change

Table 3: Descriptive statistics of data (%) of
Percentage-point Difference and Relative Percentage Change

	Amoxicillin-Free Group		**Amoxicillin-Containing Group**	
	difference	Percentage difference	difference	Percentage difference
0 vs 10	-4.9%	-5.3%	-8.1%	-9.0%
10 vs 20	-5.0%	-5.7%	-29.2%	-35.8%
20 vs 40	-12.7%	-15.3%	-2.8%	-5.3%
40 vs 80	+18.6%	+26.5%	-39.5%	-79.6%

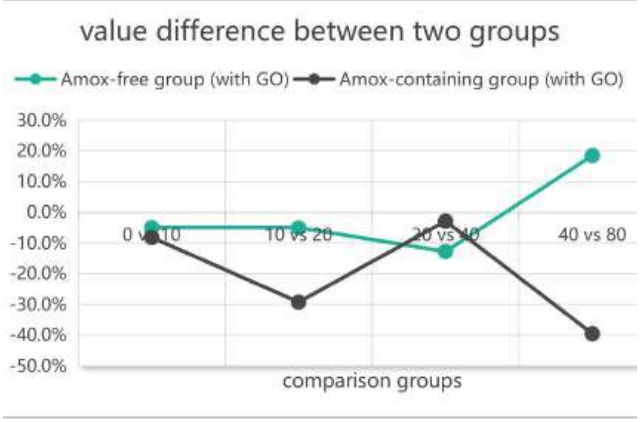


Figure 3: Value difference between two groups

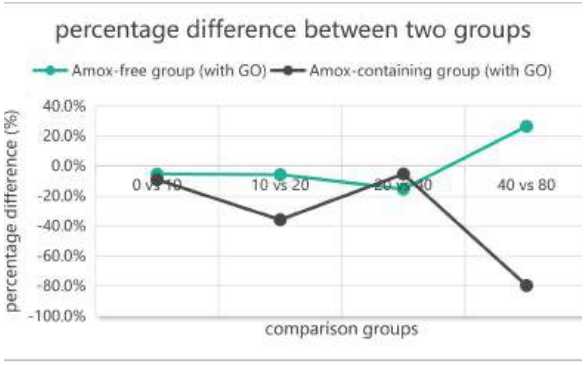


Figure 4: Percentage difference between two groups

Data Processing and Regression Analysis

1. **The formula of coefficient of determination (R^2)**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

SS_{res} (Residual Sum of Squares) $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

(i.e., the sum of squared differences between observed values and predicted values)

SS_{tot}(Total Sum of Squares) = $\sum_{i=1}^{n} (y_i - \bar{y})^2$

(i.e., the sum of squared differences between observed values and the overall mean)

2. Regression Model Selection

Based on the visual inspection of the concentration-response curves:
Amoxicillin-Containing Group: The colony area exhibited a continuous downward trend with increasing GO concentration. A linear regression model (y=a+bx) was therefore applied.
Amoxicillin-Free Group: The colony area showed a biphasic V-shaped pattern (decreasing from 0 to 40 μg/mL, then rebounding at 80 μg/mL). A second-order polynomial (quadratic) regression model (y=a+bx+cx2) was therefore applied.

3. Worked Example 1: Amoxicillin-Containing Group (Linear Model)

The linear regression equation is expressed as: y=a+bx
Where: x = GO concentration (μg/mL): 0, 10, 20, 40, 80
y = Colony area (%): 89.7, 81.6, 52.4, 49.6, 10.1
The means were calculated as:

$\bar{x} = \frac{0+10+20+40+80}{5} = 30$

$\bar{y} = \frac{89.7+81.6+52.4+49.6+10.1}{5} = 56.68$

Using the least-squares estimation formulas:

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})\ (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Table 4: detailed calculation table

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	89.7	-30	33.02	-990.6	900
10	81.6	-20	24.92	-498.4	400
20	52.4	-10	-4.28	42.8	100
40	49.6	10	-7.08	-70.8	100
80	10.1	50	-46.58	-2329.0	2500
Sum	-	-	-	$\sum = -3846.0$	$\sum = 4000$

Thus: $b = \frac{-3846.0}{4000} = -0.9615$

$$a = \bar{y} - \overline{x}b = 56.68 - (-0.9615 \times 30) = 85.525$$

The final regression equation is: y=85.525−0.9615x

4. For Amoxicillin-Free Group (Quadratic Polynomial Model)

The final regression equation is: y=95.0−1.00x+0.0107x^2

Table 5: Calculation of Residual Sum of Squares (for Amoxicillin-Containing Group)

x_i	Observed y_i	Predicted $\hat{y}_i = 85.525 - 0.9615x_i$	Residual $y_i - \hat{y}_i$	Residual² $(y_i - \hat{y}_i)^2$
0	89.7	85.525	4.175	17.43
10	81.6	76.110	5.490	30.14
20	52.4	66.695	-14.295	204.35
40	49.6	47.865	1.735	3.01
80	10.1	8.555	1.545	2.39
			$SS_{res} = \sum$	**257.32**

Table 6: Calculation of Total Sum of Squares for Amoxicillin-Containing Group

x_i	Observed y_i	$\bar{y} = 56.68$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
0	89.7	56.68	33.02	1089.12
10	81.6	56.68	24.92	620.65
20	52.4	56.68	-4.28	18.32
40	49.6	56.68	-7.08	50.12
80	10.1	56.68	-46.58	2171.19
			$SS_{tot} = \sum$	**3949.40**

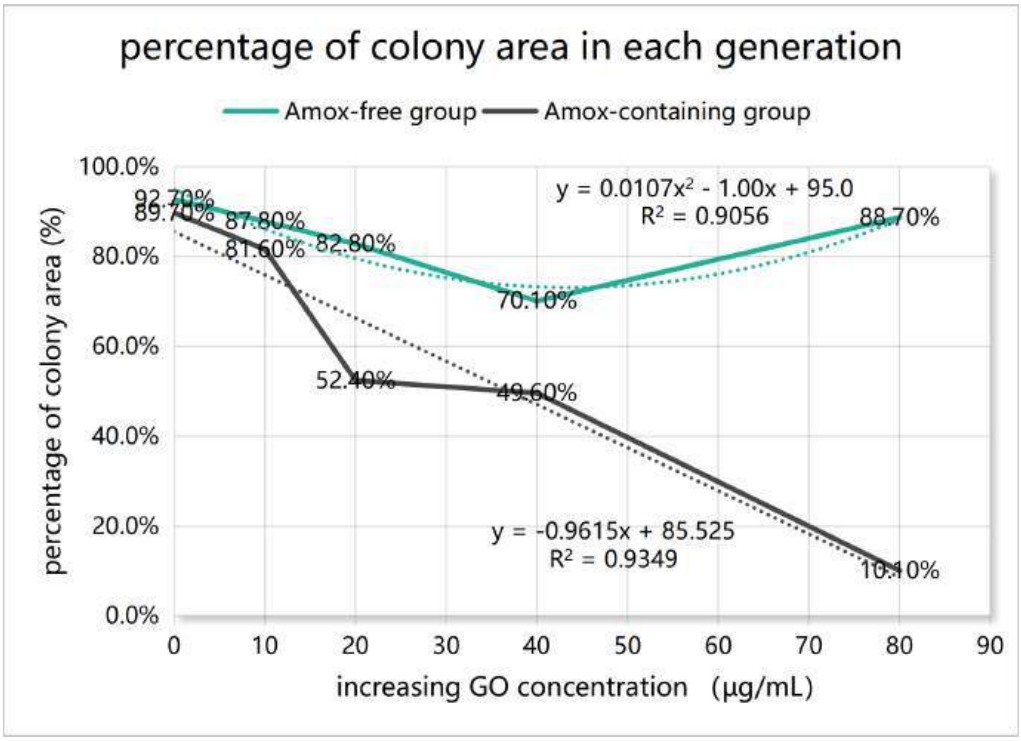
Table 7: Calculation of Coefficient of Determination for Amoxicillin-Containing Group

Parameter	Symbol	Value
Residual Sum of Squares	SS_{res}	257.32
Total Sum of Squares	SS_{tot}	3949.40
Coefficient of Determination	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$	$1 - 0.0651 = \mathbf{0.9349}$

Result: The linear model explains 93.49% of the variance in the Amoxicillin-Containing group.

5. **For Amoxicillin-Free Group**

Coefficient of Determination = 0.9056, which the linear model explains 90.56% of the variance in the Amoxicillin-Free group.



Results

Results show that graphene oxide affected the growth of *Lactobacillus acidophilus*, with different responses observed depending on the presence of amoxicillin. In the amoxicillin-free group, colony coverage decreased from 92.7% at 0 µg/mL GO to 70.1% at 40 µg/mL, before increasing to 88.7% at 80 µg/mL. In contrast, the amoxicillin-containing group showed a continuous decrease in colony coverage as GO concentration increased, from 89.7% at 0 µg/mL to 10.1% at 80 µg/mL. This indicates a stronger inhibitory effect of increasing GO concentration when amoxicillin was present.

Regarding to statistical analysis, coefficient of determination (R^2) was calculated and regression analysis was performed. Graphene oxide (GO) modulates the growth dynamics of *Lactobacillus acidophilus* in a context-dependent manner across five consecutive passages. Without antibiotics, GO exhibits a biphasic response ($R^2 = 0.9056$), a non-monotonic trend likely due to reduced effective surface contact between GO sheets and the bacterial membrane caused by nanomaterial aggregation at high concentrations. Conversely, when co-administered with 20 µg/mL amoxicillin, GO shows a significant negative linear relationship with bacterial growth ($R^2 = 0.9349$).

Discussion/Evaluation

The results indicate that graphene oxide (GO) affected the growth of *Lactobacillus acidophilus*, with the effect differing depending on the presence of amoxicillin. In the amoxicillin-free group, colony coverage decreased from 92.7% at 0 μg/mL GO to 70.1% at 40 μg/mL, before increasing to 88.7% at 80 μg/mL. In contrast, the amoxicillin-containing group showed a continuous decrease in colony coverage as GO concentration increased, reaching 10.1% at 80 μg/mL. This suggests that the growth response to GO was influenced by the presence of amoxicillin.

The increase in colony coverage at 80 μg/mL in the amoxicillin-free group may be explained by changes in the effective availability of GO, such as aggregation, which can reduce direct interaction between GO sheets and bacterial cells. Alternatively, physiological stress responses or phenotypic tolerance may have contributed to the observed rebound. However, the present experiment cannot determine whether these changes resulted from genetic adaptation or heritable antibiotic resistance because no genetic or post-exposure susceptibility testing was performed. Therefore, the increase in growth should not be interpreted as evidence that the bacteria developed antibiotic resistance.

Under antibiotic-free conditions, GO exhibited a two-sided effect: at a GO concentration of 40 μg/mL, colony coverage decreased from 92.7% to 70.1%, then rebounded to 88.7% at 80 μg/mL, likely due to nanomaterial aggregation or transient stress tolerance. In contrast, co-exposure with 20 μg/mL amoxicillin produced a strong linear inhibitory effect, with colony coverage decreasing from 89.7% to 10.1% at 80 μg/mL GO ($R^2 = 0.9349$). High concentrations of GO alone lose their antibacterial efficacy due to aggregation, but their combination with amoxicillin significantly enhances the inhibitory effect on probiotic growth.

Overall, the findings suggest that GO can alter the growth response of *L. acidophilus*, particularly under amoxicillin exposure. The results also demonstrate that increasing GO concentration does not necessarily produce the same response under different antimicrobial conditions. Further research using biological replication, longer exposure periods, and direct measurements of antibiotic susceptibility would be required to determine whether the observed changes represent physiological stress, phenotypic tolerance, or heritable adaptation.

Despite these findings, the experiment has limitations affecting its validity and generalizability. Several alternative explanations should be considered, which include the lack of biological replication, limited number of bacterial generations, uncontrolled environmental conditions, and the use of a single bacterial species and antibiotic. The observed differences in colony coverage may partly reflect technical variability, including inconsistent inoculum density, manual plating variation, or random variation. Image-threshold errors may have arisen from GO precipitation altering optical appearance, especially at high concentrations. Although no visible contamination was observed, microbiological confirmation was not performed. The growth rebound at 80 μg/mL GO may reflect nanomaterial aggregation, but this remains speculative without direct physical characterization. Addressing these limitations would improve the reliability of future investigations.

Limitations/Improvements

Limitation: Only technical replication was performed instead of biological replication,so the experiment results may lack ecological validity, statistical inference and generalization to population-level responses are limited.

Improvement: Repeat the experiment with independently cultured bacterial colonies across three separate days to establish biological replication.

Limitation: The bacteria was planted in agar without the use of a pipette, which may result in systemic errors.

Improvement: Use a micropipette with sterile tips to transfer a standardized volume (e.g., 100 µL) of bacterial suspension onto each agar plate, ensuring uniform inoculation.

Limitation: This experiment cannot be generalized onto other species of bacteria, antibiotics, nor nanomaterials.

Improvement: Expand the experimental design to include additional bacterial strains (e.g., *E. coli, B. subtilis*), other antibiotics (e.g., tetracycline), and different nanomaterials (e.g., reduced graphene oxide, carbon nanotubes) to assess broader applicability.

Limitation: The experiment was conducted over a relatively short span of time so assumptions cannot be made about bacterial resistance over further generations

Improvement: Extend the subculturing period to 15–30 generations and periodically store bacterial samples for post-experiment MIC testing to evaluate long-term adaptive responses.

Limitation: The petri dishes were not kept under strict conditions of 37 degrees Celsius nor strict anaerobic.aerobic conditions, which may also interfere with the accuracy of the results to unknown extents

Improvement: Use a temperature-controlled incubator set to 37°C and an anaerobic jar with gas-generating sachets to maintain consistent growth conditions across all plates.

Limitation: The presence of unpredictable fluctuations in measurement that can cause low precision, and potential systematic errors involving faulty procedures or errors regarding materials (purity, environmental factors, etc.)

Improvement: Averaging multiple repeat experiments to reduce random errors, removing zero errors, and eliminating as many confounding variables as possible.

Acknowledgements

The authors would like to thank the University of Malta's Department of Metallurgy and Materials Engineering for kindly providing the graphene oxide used in this study.

We are also grateful to Dr. Ing. Anthea Agius Anastasi and Ms. Radka Zarkova for their valuable guidance and feedback throughout this work.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Works Cited

Automeris.io. (n.d.). Web-based image analysis for colony area quantification. Retrieved March 30, 2026, From https://automeris.io

Akhavan, B. J., Vijhani, P., & Khanna, N. (2023, November 17). Amoxicillin. Retrieved from National Library of Medicine website: https://www.ncbi.nlm.nih.gov/books/NBK482250/

Anegbe, B., Ifijen, I., Maliki, M., Uwidia, I., & Aigbodion, A. (2024). Graphene oxide synthesis and applications in emerging contaminant removal: a comprehensive review. Environmental Sciences Europe, 36, 1-34. https://doi.org/10.1186/s12302-023-00814-4.

Gao, H., Li, X., Chen, X., Hai, D., Wei, C., Zhang, L., & Li, P. (2022). The Functional Roles of Lactobacillus acidophilus in Different Physiological and Pathological Processes. Journal of Microbiology and Biotechnology, 32(10), 1226–1233. https://doi.org/10.4014/jmb.2205.05041

Guo, M., & Zhang, G. (2017). Graphene oxide in the water environment could affect tetracycline-antibiotic resistance.. Chemosphere, 183, 197-203 . https://doi.org/10.1016/j.chemosphere.2017.04.145.

Jiříčková, A., Jankovský, O., Sofer, Z., & Sedmidubský, D. (2022). Synthesis and Applications of Graphene Oxide. Materials, 15(3), 920. https://doi.org/10.3390/ma15030920

Liu, S., Zeng, T., Hofmann, M., Burcombe, E., Wei, J., Jiang, R., Kong, J., & Chen, Y. (2011). Antibacterial activity of graphite, graphite oxide, graphene oxide, and reduced graphene oxide: membrane and oxidative stress.. ACS nano, 5 9, 6971-80 . https://doi.org/10.1021/nn202451x.

Montes-Duarte, G., Tostado-Blázquez, G., Castro, K., Araujo, J., Achete, C., Sánchez-Salas, J., & Campos-Delgado, J. (2021). Key parameters to enhance the antibacterial effect of graphene oxide in solution. RSC Advances, 11, 6509 - 6516. https://doi.org/10.1039/d0ra07945f.

Pulingam, T., Thong, K., Appaturi, J., Nordin, N., Dinshaw, I., Lai, C., & Leo, B. (2020). Synergistic Antibacterial Actions of Graphene Oxide and Antibiotics towards Bacteria and the Toxicological Effects of Graphene Oxide on Human Epidermal Keratinocytes.. European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences, 105087 . https://doi.org/10.1016/j.ejps.2019.105087.

Simeonova, D. D., Pollmann, K., Bianco, A., & Didier Lièvremont. (2024). Graphene oxide and bacteria interactions: what is known and what should we expect? MSphere. https://doi.org/10.1128/msphere.00715-23

Witsell, D. L., Garrett, C. G., Yarbrough, W. G., Dorrestein, S. P., Drake, A. F., & Weissler, M. C. (1995). Effect of Lactobacillus acidophilus on antibiotic-associated gastrointestinal morbidity: a prospective randomized trial. The Journal of Otolaryngology, 24(4), 230–233. Retrieved from https://pubmed.ncbi.nlm.nih.gov/8551535/

Zakhar Ivanovich Evseev, Lidia Andreevna Tarasova, Fedora Dmitrievna Vasilieva, Marfa Nikitichna Egorova, Petr Stanislavovich Dmitriev, Yana Aleksandrovna Akhremenko, & Svetlana Afanasyevna Smagulova. (2024). Comparison of Antimicrobial Properties of

Graphene Oxide-Based Materials, Carbon Dots, and Their Combinations Deposited on Cotton Fabrics. International Journal of Molecular Sciences, 25(10), 5328–5328. https://doi.org/10.3390/ijms25105328

Zhang, Q., & Zhang, C. (2020). Chronic Exposure to Low Concentrations of Graphene Oxide Increases Bacterial Pathogenicity via the Envelope Stress Response.. Environmental science & technology. https://doi.org/10.1021/acs.est.0c04538

Zhang, T., & Tremblay, P.-L. (2020). Graphene: An Antibacterial Agent or a Promoter of Bacterial Proliferation? IScience, 23(12), 101787. https://doi.org/10.1016/j.isci.2020.101787

Domestic Stability, International Distrust: The FFKR's Post-2022 Brand Rehabilitation
By Maria Runje

Section 1. Introduction

1.1 Background

The Beijing Winter Olympics in February 2022 laid the groundwork for one of the most destructive institutional crises in the history of modern figure skating. The star of the Russian Olympic Committee's figure skating team, Kamila Valieva, then fifteen, was found to have failed a drug test for trimetazidine, a banned heart medication, in a sample taken at the Russian National Championships in December 2021. The Court of Arbitration for Sport (CAS) still let her compete in Beijing following the positive test, because her status as a minor brought special procedural protections. The scandal did not dissipate quickly. After a lengthy legal battle involving competing appeals from the World Anti-Doping Agency (WADA), the International Skating Union (ISU) and Russia's own anti-doping agency (RUSADA), the CAS issued its decision in January 2024. Valieva was handed a four-year ban, and had all of her results from December 2021 nullified, including the Russian team's gold medal in the Beijing team event (Mann). The United States, originally awarded silver, was upgraded to gold at a special ceremony at the Paris 2024 Summer Olympics (International Skating Union, "Final Ranking Confirmed").

The Figure Skating Federation of Russia (FFKR)'s reputational crisis did not only stem from the doping case. Four days after the closing ceremony of the Beijing Games, Russia launched a full-scale military invasion of Ukraine. On March 1, 2022, the ISU issued a ban on all Russian and Belarusian figure skaters and officials from participating in international competitions "to protect the integrity of ice skating competitions" (International Skating Union, "ISU Statement on the Ukrainian Crisis"). This ban was then extended for the 2022-23, 2023-24 and 2024-25 seasons, and, in 2024, the ISU announced that the ban would remain in place indefinitely for ISU sanctioned events (International Skating Union, "ISU Statement on the Ukrainian Crisis"; International Skating Union, "ISU Decision"). The FFKR found itself facing an acute, multidimensional reputational crisis, exacerbated by a high-profile doping scandal involving a minor, prolonged international exclusion, and intense scrutiny of its coaching culture. Particular attention focused on Eteri Tutberidze, Valieva's coach, who was publicly criticised by IOC President Thomas Bach for her treatment of the fifteen-year-old following the competition ("'Chilling'").

From a business management perspective, the FFKR is an organisational entity whose "brand image," defined here as the sum of the perceptions and associations held by its key stakeholder groups, is central to its operational viability and institutional legitimacy. Academic research demonstrates that anti-doping violations can seriously affect sports organisations, damaging consumer trust, spectator attendance, and sponsorship relationships in the aftermath of a scandal ((Chien et al. 498; Sassenberg and Sassenberg 1044). Where such violations involve a

minor athlete and potential institutional negligence, the reputational damage extends beyond the individual athlete to the organisation itself. From 2023 onward, the FFKR therefore faced a dual challenge: managing the immediate crisis while beginning the longer-term process of brand reconstruction.

This essay analyses the effectiveness of the FFKR's crisis-management strategy between 2023 and 2026 in rehabilitating its brand image. First, it considers the nature and severity of the crisis through stakeholder analysis and Coombs' Situational Crisis Communication Theory. Second, it examines the FFKR's strategic response through a SWOT analysis, focusing on its domestic competition strategy, institutional communications, and approach to coaching reform. Third, it evaluates the effects of these measures on domestic and international stakeholders using force-field analysis and the CUEGIS concepts of ethics and globalisation. The essay argues that, although the FFKR has achieved meaningful domestic brand stabilisation, its international brand image remains severely constrained, not primarily because of strategic failure, but because of structural and geopolitical barriers beyond the federation's direct control.

Section 2: Methodology

Part 1: Resources

This essay draws primarily on secondary research, supplemented by official primary documents. The sources can be divided into four categories: official and institutional documents, international news media, Russian state media, and peer-reviewed academic literature.

The most authoritative evidence comes from official and institutional sources. The International Skating Union's statement on the Individual Neutral Athlete pathway for the 2026 Milan Cortina Winter Olympics provides direct documentation of the conditions under which Russian athletes were permitted to re-enter international competition on a limited basis (International Skating Union, "ISU Decision"). The Court of Arbitration for Sport ruling concerning Kamila Valieva's suspension provides the principal legal evidence for the crisis timeline. Where the full ruling is not directly cited, reporting by NPR and PBS NewsHour is used to explain its findings and implications (Mann; Dunbar). These sources are considered highly credible because their claims are institutionally attributable and independently verifiable.

International news organisations, including NPR, CBC Sports, NBC News, Inside the Games, and the Associated Press, provide contemporaneous coverage of significant events between 2022 and 2026. Although the editorial focus of each organisation differs, these sources are generally regarded as credible because of their established reporting standards. Where possible, major claims are compared across multiple outlets to reduce reliance on a single interpretation.

Russian state media, particularly TASS, is used sparingly and for a specific analytical purpose. The closest available primary evidence of the federation's position is a May 2025 statement made by FFKR President Anton Sikharulidze to TASS concerning the approval of neutral athletes for the 2026 Olympic qualifying process (TASS). However, TASS is a Russian

state news agency, and its reporting on institutional responses may present the FFKR in a favourable manner. It is therefore used primarily to analyse the federation's communication strategy and preferred public narrative rather than as independent evidence of the effectiveness of its actions.

The analytical framework is derived from peer-reviewed academic literature. The classification of the crisis and the assessment of stakeholder responses are based on Coombs and Holladay's development of Situational Crisis Communication Theory and Brown et al.'s application of SCCT to sports organisations (210). Research by Chien et al. and Sassenberg and Sassenberg on consumer and sponsor reactions following athlete scandals provides the foundation for evaluating the commercial and reputational dimensions of the crisis. These studies constitute the most methodologically rigorous part of the source base because they have undergone peer review and use established research methods.

The principal limitation of the investigation is the restricted availability of direct primary evidence. Geographic, linguistic, and institutional barriers make it difficult to obtain testimony from Russian figure-skating stakeholders, domestic audiences, coaches, and FFKR officials. The assessment of domestic brand image consequently relies largely on indirect indicators, including competition activity, media coverage, audience engagement, and observable institutional behaviour, rather than direct surveys of stakeholder perceptions. This limitation is acknowledged throughout the analysis. The conclusions should therefore be understood as evidence-informed judgements rather than definitive measurements of the FFKR's brand image.

Part 2: Tools for Analysis

This essay applies four business management tools, each selected for its relevance to a specific dimension of the crisis.

Section 3 applies Coombs and Holladay's Situational Crisis Communication Theory (SCCT) to classify the crisis affecting the FFKR and assess the resulting level of reputational threat. SCCT distinguishes among victim, accidental, and preventable crises. Because the scandal involved a minor athlete within an institutional training system and raised questions about organisational responsibility, the FFKR's crisis can be classified as preventable. According to SCCT, this category generates the highest level of reputational threat because stakeholders attribute substantial responsibility to the organisation. This classification provides the starting point for evaluating the federation's subsequent brand-rehabilitation efforts.

Stakeholder analysis is applied alongside SCCT to identify the FFKR's principal stakeholder groups, including domestic audiences, international governing bodies such as the ISU, WADA, and IOC, sponsors, athletes, coaches, and federation officials. It is then used to assess how the crisis and the FFKR's response affected each group differently. This distinction is central to the essay's argument that the federation's domestic and international brand-rehabilitation efforts have produced significantly different outcomes.

Section 4 uses a SWOT analysis to evaluate the FFKR's strategic position from 2023 onward. The analysis identifies internal strengths, including the depth of its domestic talent

pipeline, supportive domestic media coverage, and an established national competition infrastructure. It also examines weaknesses such as the loss of an international platform and concerns about athlete welfare and institutional credibility. External opportunities include the possibility of a limited return to international competition through the Individual Neutral Athlete pathway, while threats include continued ISU restrictions, geopolitical isolation, and sustained scrutiny of the federation's coaching culture. SWOT is particularly relevant because it distinguishes between internal factors that the FFKR can influence and external conditions that remain largely beyond its control.

Section 5 uses force-field analysis to evaluate the effectiveness of the FFKR's brand-rebuilding efforts. The analysis compares the forces supporting reputational recovery, including domestic state support, continuity in the talent pipeline, and the limited ISU pathway for re-entry, with the forces restricting international rehabilitation, such as the indefinite ISU ban, continued WADA scrutiny, the controversy surrounding Eteri Tutberidze, and the wider geopolitical context. This tool is particularly useful because it helps explain why the federation's crisis-management strategy has produced asymmetric outcomes among domestic and international audiences, despite a degree of strategic coherence.

The CUEGIS concepts of ethics and globalisation are integrated throughout Sections 4 and 5 to provide a broader analytical framework. The ethical dimension concerns the unresolved question of institutional responsibility for the doping of a minor and whether the FFKR's response represented genuine structural reform or primarily cosmetic reputation management. The concept of globalisation places the crisis within the broader international isolation of Russian sport and examines the extent to which a national federation can rebuild its brand while excluded from global competitive and institutional structures.

Section 3: Nature and Severity of the Crisis

3.1. Situational Crisis Communication Theory: Crisis Classification

Coombs and Holladay's Situational Crisis Communication Theory (SCCT) provides an effective framework for evaluating the FFKR's crisis-management strategy. The theory classifies crises into three broad clusters—victim, accidental, and preventable—according to the level of responsibility attributed to the organisation. Preventable crises, in which an organisation is perceived to have knowingly engaged in harmful or inappropriate conduct, create the greatest reputational threat and therefore require the most substantial corrective response (Coombs and Holladay 175). In the context of the FFKR, SCCT is used to classify the nature and severity of the crisis and to assess whether the federation's response was proportionate to the level of responsibility attributed to it.

The FFKR's crisis can reasonably be classified within SCCT's preventable cluster because stakeholders were likely to attribute responsibility not only to Valieva, but also to the institutional system responsible for supervising and protecting a minor athlete. In January 2024, CAS found that Valieva had committed an anti-doping rule violation after trimetazidine was

detected in a sample collected at the Russian National Championships on 25 December 2021. CAS imposed a four-year period of ineligibility and disqualified her competitive results from that date onward (Court of Arbitration for Sport).

Although the ruling formally concerned Valieva's individual responsibility, she was only fifteen at the time, causing international scrutiny to extend to the coaches and institutions surrounding her. This scrutiny intensified after her free skate at the 2022 Beijing Olympics, when IOC President Thomas Bach criticised the "tremendous coldness" shown by her entourage and described its treatment of her as "chilling" ("'Chilling'"). The incident therefore reinforced perceptions that the crisis reflected wider concerns about athlete welfare and the FFKR's coaching culture rather than the isolated misconduct of one athlete.

However, Tutberidze was not found to have participated in the doping process and was not formally sanctioned. Her return to the 2026 Milan Cortina Olympics as the accredited coach of Georgian skater Nika Egadze nevertheless renewed controversy. WADA President Witold Bańka stated that he felt uncomfortable with her presence, while acknowledging that the investigation had found no legal basis for excluding her (Grohmann). This combination of continued public concern and limited institutional accountability increased the reputational threat facing the FFKR, supporting its classification as a preventable crisis from a stakeholder-perception perspective.

Two additional factors intensified the reputational threat beyond the initial SCCT classification. First, the FFKR faced a significant crisis in history. Russian sporting institutions had previously been associated with state-sponsored doping, most notably through the McLaren Report (2016), which documented systematic manipulation of doping controls across Russian sport, including during the 2014 Sochi Winter Olympics. This history made international stakeholders more likely to interpret the Valieva case not as an isolated act of individual misconduct, but as evidence of a broader institutional pattern.

Second, the institutional response lacked a clear, unified, and credible explanation capable of restoring stakeholder confidence. RUSADA's appeal to CAS sought a reprimand for Valieva rather than her complete exoneration (Nivison), which weakened attempts to portray the incident as entirely unfounded. However, because RUSADA is institutionally separate from the FFKR, this should be treated as contextual evidence of the wider Russian sporting response rather than direct proof of collusion by the federation. Under SCCT, a negative crisis history and an unfavourable prior reputation increase stakeholder attribution of responsibility and magnify reputational damage (Coombs and Holladay 175). The FFKR therefore faced not only the immediate consequences of a preventable crisis, but also the cumulative effects of longstanding international distrust.

3.2 Stakeholders Analysis

The crisis and the federation's response were experienced in fundamentally different ways across stakeholder groups. A stakeholder analysis identifies four principal groups, each with distinct interests, levels of influence, and perceptions of the FFKR.

International governing bodies (ISU , WADA , IOC) represent the most powerful and negatively impacted stakeholder group. The ISU's ban on Russian and Belarusian skaters, first implemented on 1 March 2022, and extended indefinitely in 2024, effectively removed the athletes of the FFKR from the international competitive landscape (International Skating Union, "ISU Statement on the Ukrainian Crisis"; International Skating Union, "ISU Decision"). For this stakeholder group, the FFKR's brand image was shaped not only by the doping scandal but also by the geopolitical context of Russia's invasion of Ukraine — an element that lies beyond the federation's control. The restrictions placed on Russian athletes applying for neutral status for the 2026 Milan Cortina Olympics – intensive political vetting, exclusion from team events, a ban on media engagement that was later partially rescinded, and no display of the Russian flag or anthem – reveal just how much institutional mistrust is directed at Russian sport in its entirety (International Skating Union, "ISU Decision"; International Skating Union, "Special Screening Process"; Zaccardi). Critically, only four Russian figure skaters passed the ISU's screening process for the 2026 qualifiers and none were approved in pairs or ice dance disciplines (International Skating Union, "Special Screening Process"; Zaccardi). This shows that the federation's brand rehabilitation in this stakeholder group was still severely limited as of 2025.

The stakeholder dynamic was markedly different among domestic audiences and state media. Russian outlets, including TASS and Match TV, frequently framed the Valieva case as a politically motivated attack on Russian sport by Western institutions rather than as a legitimate anti-doping intervention (TASS). This narrative appears to have helped protect the FFKR's domestic reputation by redirecting blame toward external actors.

The federation also maintained a strong domestic competition infrastructure throughout the crisis. This is reflected in the continued expansion and promotion of national events, including the renaming of the domestic cup series as the "Russian Grand Prix" for the 2022 to 2023 season and the continued staging of the Russian Figure Skating Championships in major cities such as Krasnoyarsk, Chelyabinsk, and Saint Petersburg (Figure Skating Federation of Russia, "Grand Prix of Russia Idel-2022"; Figure Skating Federation of Russia, "Russian Championship 2023"; Figure Skating Federation of Russia, "Russian Championship 2024"; Figure Skating Federation of Russia, "Russian Championship 2026"). These developments suggest that, although the FFKR experienced severe international reputational damage, its domestic brand remained comparatively stable among this stakeholder group.

The most difficult stakeholder group to assess is commercial sponsors because publicly available information about the FFKR's specific sponsorship relationships remains limited. Research into consumer responses to sponsorship following athlete scandals suggests that sponsor brands can experience measurable reputational spillover when associated with doping violations, particularly when the case involves a minor athlete (Chien et al. 498). Any international sponsors connected to the FFKR before 2022 would therefore have faced considerable pressure to distance themselves from Russian sport following both the Valieva scandal and Russia's invasion of Ukraine. However, the federation's shift toward a domestically oriented model from 2023 onward, supported by continued exposure through Russian state

media, suggests a strategic repositioning away from international commercial relationships and toward domestic and state-backed sponsorship structures. This represented a rational response to the loss of international commercial viability, although it also reinforced the growing divergence between the FFKR's domestic and international brand trajectories.

The final key stakeholder group consists of athletes and coaching staff. Tutberidze's continued presence within the FFKR's competitive system, including through her student Adeliia Petrosian, who competed as a neutral athlete at Milano Cortina 2026 and was identified as a potential gold-medal contender, suggests that the federation made limited effort to distance itself structurally from the coaching model associated with the Valieva scandal (Dedaj; "Who Is Adeliia Petrosian"). Although this continuity may have supported competitive success, it also reinforced international concerns about athlete welfare and institutional accountability. The decision therefore had significant implications for the ethical dimension of the FFKR's brand image, which is examined further in Section 5.

3.3 Summary

SCCT classifies the FFKR's crisis as preventable, representing the highest level of reputational threat within the theory's framework. This threat was intensified by Russia's previous doping history and the absence of a clear, credible, and corrective institutional response. The stakeholder analysis reveals a fundamental asymmetry. Domestic audiences and state-aligned stakeholders experienced a contained and narratively managed crisis, while international governing bodies and commercial stakeholders encountered a sustained crisis reinforced by institutional distrust and geopolitical exclusion. This divergence between a comparatively stable domestic brand and a severely damaged international reputation provides the central analytical thread for the remainder of the essay.

Section 4: The FFKR's Crisis Management Strategy A SWOT Analysis

After defining the nature and extent of the crisis in Section 3, this section assesses the FFKR's strategic response from 2023-2026. A SWOT analysis provided the structural framework for identifying the internal strengths and weaknesses the federation brought to its crisis management and the external opportunities and threats which shaped its strategic options. This analysis provides a direct answer to the RQ by studying not only the actions of the FFKR, but also an evaluation of whether its choices were strategically consistent given its position.

4.1 Strengths

The FFKR entered the post-scandal period with two significant internal strengths that partially supported its capacity to manage the crisis. First, it possessed a deep pool of domestic talent. Despite the international restrictions, Russian figure skating retained several world-class athletes, including Adeliia Petrosian, Evgeni Semenenko, and Petr Gumennik, which helped preserve the quality of domestic competition and public interest in the sport (International Skating Union, "ISU Statement on the Ukrainian Crisis"; International Skating Union, "ISU

Decision"). This depth of talent enabled the federation to reframe its domestic programme not as a diminished substitute for international participation, but as a credible and largely self-sufficient competitive ecosystem.

Second, the FFKR benefited from considerable support from state media. Russian outlets such as Match TV and TASS consistently framed the Valieva case as political persecution by Western institutions rather than as a legitimate anti-doping enforcement outcome (TASS). This narrative environment reduced the need for the federation to communicate directly with domestic audiences about the crisis, as much of the reputational defence was carried out by the state media apparatus on its behalf. Although this contributed to the stability of the FFKR's domestic brand, it also revealed a structural dependence on state-controlled narratives, creating potential long-term risks if public or political support were to weaken.

4.2 Weaknesses

The federation's most significant internal weakness was the apparent absence of substantial structural reform following the scandal. Eteri Tutberidze's treatment of Valieva had been publicly criticised by IOC President Thomas Bach, yet the wider coaching culture associated with her remained largely unaddressed throughout the crisis period. Tutberidze continued to operate within the FFKR's competitive ecosystem, and her student Adeliia Petrosian later competed as a neutral athlete at the 2026 Milan Cortina Olympics (Dedaj). Furthermore, the sources reviewed provide little public evidence of changes to athlete welfare protocols, minimum-age policies, or anti-doping oversight structures during the years following the scandal. This lack of visible reform weakened the federation's credibility among international stakeholders seeking evidence that it had addressed the institutional conditions surrounding the crisis. Brown et al. argue that organisations facing preventable crises must demonstrate corrective action to begin rebuilding their reputations, as apologies and repositioning strategies have limited credibility without substantive institutional change.

A second weakness was the federation's limited ability to communicate directly with international audiences. The exclusion of Russian athletes and officials from ISU competitions reduced the FFKR's visibility within international figure skating and restricted its access to the platforms through which it might have addressed governing bodies, journalists, and other stakeholders directly (International Skating Union, "ISU Decision"). Its external messaging was therefore either channelled through Russian state media, which reinforced domestic narratives but lacked credibility among many international audiences, or confined to occasional statements by FFKR President Anton Sikharulidze through TASS, whose reach and influence were largely concentrated within Russian-language media environments (TASS). As a result, the federation struggled to present a credible and independently verifiable reform narrative to international stakeholders.

4.3 Opportunities

The most significant external opportunity available to the FFKR was the ISU's decision

in December 2024 to establish a limited neutral-athlete pathway for the 2026 Milan Cortina Olympics (International Skating Union, "ISU Decision"). This represented the first formal mechanism through which Russian skaters could return to international competition since March 2022. The FFKR responded quickly by nominating athletes for the screening process, with federation president Anton Sikharulidze confirming Adeliia Petrosian and Petr Gumennik as its leading candidates (TASS). Both athletes ultimately qualified and competed at Milan Cortina as neutral athletes, with Petrosian identified as a potential gold-medal contender ("Who Is Adeliia Petrosian"). Although highly restricted, this partial return provided the federation with a tangible opportunity to demonstrate the continuing strength of Russian figure skating on an international stage. It therefore offered a limited means of rehabilitating the sport's athletic reputation, even without the use of Russian national symbols.

The domestic competition infrastructure also represented an important strategic opportunity. The FFKR's decision to rebrand its domestic cup series as the "Russian Grand Prix" from the 2022 to 2023 season, together with its continued hosting of the Russian Championships in major cities across the country, demonstrated its ability to maintain institutional momentum and public engagement independently of international competition structures (Figure Skating Federation of Russia, "Grand Prix of Russia Idel-2022"; Figure Skating Federation of Russia, "Russian Championship 2023"; Figure Skating Federation of Russia, "Russian Championship 2024"; Figure Skating Federation of Russia, "Russian Championship 2026"). This domestic platform allowed the federation to preserve athlete visibility, spectator interest, and competitive continuity while international access remained restricted.

4.4 Threats

Despite the efforts undertaken by the FFKR, the indefinite ban imposed by the ISU presents a significant threat to the federation's future brand rehabilitation. Regardless of the FFKR's internal actions, continued exclusion from international competition restricts its ability to demonstrate reform, rebuild stakeholder trust, and restore its reputation among global audiences. The ban therefore threatens the organisation's capacity to repair its international image by removing access to the competitive platforms that generate global visibility, sponsorship value, and institutional credibility (International Skating Union, "ISU Statement on the Ukrainian Crisis"; International Skating Union, "ISU Decision").

Unlike a time-limited suspension, an indefinite ban also creates significant uncertainty and makes long-term strategic planning extremely difficult. Because the conflict in Ukraine continued without resolution throughout the 2023 to 2026 period, the geopolitical conditions required for broader international reintegration remained absent. A further threat arose from Tutberidze's continued prominence within Russian figure skating and WADA's public unease regarding her presence at Milan Cortina 2026 (Grohmann). This ensured that concerns surrounding the FFKR's coaching culture and institutional accountability remained active within international public discourse. Research on post-scandal sponsorship behaviour suggests that rebuilding consumer and stakeholder trust is particularly difficult when individuals associated

with a crisis remain in prominent positions (Chien et al. 498). This threat was therefore not merely symbolic, but carried significant implications for the FFKR's long-term international brand rehabilitation.

4.5 Strategy Analysis

The SWOT analysis reveals a crisis management strategy that was tactically effective in the domestic sphere but structurally limited in its ability to address international reputational damage. By relying on supportive state media, a deep domestic talent pool, and an established national competition infrastructure, the FFKR maintained public engagement and stabilised its brand among domestic stakeholders. However, these strengths did little to rebuild credibility among international audiences.

The federation's failure to implement visible athlete welfare and institutional reforms represented a significant strategic weakness that undermined the credibility of its international repositioning efforts. In SCCT terms, the FFKR relied primarily on bolstering and minimisation strategies, emphasising athletic excellence and portraying the crisis as politically motivated. However, it did not demonstrate the corrective action typically expected following a preventable crisis, such as visible policy reform, stronger oversight, or changes to the coaching environment (Brown et al. 210). Section 5 examines how this strategic gap contributed to the contrasting outcomes of domestic brand stabilisation and continued international reputational damage.

Section 5: Effectiveness Assessment – Force Field Analysis and CUEGIS

5.1 Analysis of Force Field

Force field analysis compares the forces supporting the FFKR's brand rehabilitation with those restricting it. This framework is particularly relevant because the balance of forces differs substantially between domestic and international audiences.

The strongest driving force was the continuity of elite athletic talent. Petrosian's status as a potential gold medal contender at Milan Cortina 2026, together with Gumennik's qualification, demonstrated that Russian figure skating remained competitively credible despite several years of exclusion from international events ("Who Is Adeliia Petrosian"; TASS). Athletic excellence represents a fundamental source of brand equity for a sporting federation. The international performances of Petrosian and Gumennik therefore provided the FFKR with a limited opportunity to demonstrate competitive strength and partially offset the reputational damage associated with the preceding crisis.

A second driving force was the continued strength of the domestic competition infrastructure. The domestic cup series was rebranded as the "Russian Grand Prix" for the 2022 to 2023 season, while the Russian Championships continued to be staged in major cities. This demonstrated the federation's ability to maintain competitive activity, athlete visibility, and institutional momentum independently of international structures (Figure Skating Federation of Russia, "Grand Prix of Russia Idel-2022"; Figure Skating Federation of Russia, "Russian

Championship 2023"; Figure Skating Federation of Russia, "Russian Championship 2024"; Figure Skating Federation of Russia, "Russian Championship 2026").

State media support also protected the FFKR's domestic brand. By framing the Valieva case as political persecution by Western institutions rather than as legitimate anti-doping enforcement, state-aligned media redirected responsibility toward external actors and reduced the domestic reputational consequences for the federation (TASS). Finally, the ISU's neutral-athlete pathway for Milan Cortina 2026 provided a limited but significant mechanism for partial international reintegration. The FFKR was able to convert this opportunity into a reputational asset by successfully fielding two qualified athletes (International Skating Union, "ISU Decision"; International Skating Union, "Special Screening Process"; Zaccardi).

However, the restraining forces were substantially more powerful, particularly at the international level. The ISU's continuing restrictions removed access to the competitive platforms that generate global visibility, sponsorship value, and institutional legitimacy. Without a geopolitical resolution to the conflict in Ukraine, this constraint remained largely beyond the federation's control throughout the period examined (International Skating Union, "ISU Statement on the Ukrainian Crisis"; International Skating Union, "ISU Decision").

A second major restraining force was Tutberidze's continued prominence within the competitive system. WADA's public unease regarding her presence at Milan Cortina 2026 demonstrated that concerns surrounding her methods and the wider coaching culture remained unresolved (Grohmann). Research suggests that stakeholder trust is especially difficult to rebuild when individuals associated with the original crisis continue to hold prominent positions (Chien et al. 498). Russia's broader history of doping represented a further constraint. The findings of the McLaren Report (2016) created a negative crisis history that encouraged international stakeholders to interpret the Valieva case as evidence of a systemic problem rather than an isolated incident. Under SCCT, such a crisis history increases the level of responsibility attributed to an organisation and intensifies reputational damage (Coombs and Holladay 175). This made the FFKR's international reputation difficult to rehabilitate through communications or competitive performance alone.

The overall balance of forces is therefore highly uneven. Most of the driving forces operated within the domestic environment, while the strongest restraining forces affected the federation internationally. The FFKR's strategy may have been internally coherent and effective in preserving domestic legitimacy, but domestic stability did not translate into international reputational recovery.

5.2 CUEGIS: Globalisation and Ethics

The ethical dimension of the FFKR's crisis ultimately concerns whether its response represented genuine institutional reform or primarily reputation management. The available evidence weighs more heavily toward the latter. Between 2023 and 2026, there was no publicly documented announcement of substantial changes to athlete-welfare protocols, and no members of the coaching staff were formally sanctioned. Tutberidze continued to operate without

institutional repercussions, while the FFKR's selection of her student, Petrosian, as its leading candidate for neutral participation could be interpreted as an implicit endorsement of a training model whose ethical standards had been publicly questioned ("Who Is Adeliia Petrosian"). Brown et al. demonstrate that, in preventable crises, organisations that fail to implement visible corrective action risk undermining the credibility of the wider crisis response (210). The FFKR's limited evidence of structural reform therefore compounded its international reputational difficulties, even if geopolitical exclusion remained the most significant external barrier to rehabilitation.

The globalisation dimension reveals the structural limits on what any crisis management strategy could realistically achieve. The ISU controls access to the Grand Prix series, World Championships, and Olympic Games, which are the principal stages on which international brand equity in figure skating is built. From March 2022 onward, exclusion from these competitions effectively disconnected the FFKR from the global infrastructure on which international brand rehabilitation depends.

This does not excuse the federation's strategic shortcomings, but it does require the conventional SCCT framework to be qualified. Where the principal mechanisms of reputational recovery are geopolitically inaccessible, the ceiling on rehabilitation becomes structural rather than purely strategic. The sharp divergence between the FFKR's domestic and international brand images is therefore also a product of geopolitical conditions. For domestic audiences, supported by state media narratives, the federation may remain associated with athletic excellence and national resilience. For international audiences, it is more closely associated with doping, institutional evasion, and a damaging coaching culture.

These represent, in effect, two distinct brand realities produced by the same organisation. This bifurcation could not be overcome through domestic strategy alone, regardless of how effectively that strategy was implemented.

Section 6: Conclusion

This essay argued that the FFKR's crisis management strategy from 2023 to 2026 stabilized the domestic brand in meaningful ways, but failed to achieve substantial international rehabilitation. From the outset, the FFKR was a severe and multidimensional reputational challenge to the SCCT, as it was a preventable crisis (the highest reputational threat category) that was compounded by Russia's past doping record and the lack of a credible institutional denial. Its strategic response was tactically competent in the domestic sphere, mobilising talent depth, state media support and a rebranded domestic competition infrastructure to keep the public engaged and institutional momentum alive. However, on a global scale, the federation's inability to execute any tangible structural reform—most importantly, its continued affiliation with Tutberidze's coaching model—tarnished the credibility of its repositioning campaign with the stakeholder groups whose trust it most desperately needed to regain.

The force field and CUEGIS analyses combined reveal that this international failure was only partly the result of strategic choices. An indefinite ISU ban, geopolitically conditioned and

structurally immovable within the timeframe of this essay, imposed a ceiling on international brand rehabilitation that no crisis management strategy could have fully overcome. Partial re-entry via the Milan Cortina 2026 neutral athlete pathway showed the ceiling was not absolute, but with two athletes competing without a flag and the ban still in place recovery remained severely constrained. Ultimately, the case of the FFKR illustrates a fundamental limit of crisis management theory as it is applied to organisations that operate within – or, in this case, outside of – the structures of globalised sport: institutional reputation cannot be rebuilt on platforms to which the institution has no access.

Works Cited

Brown, Kenon A., et al. “Applying Situational Crisis Communication Theory to Sports: Investigating the Impact of Athlete Reputational Crises on Team Perception.” Journal of Global Sport Management, vol. 5, no. 2, 2020, pp. 202–222. https://doi.org/10.1080/24704067.2019.1604077.

Chien, P. Monica, et al. “Sport Scandal and Sponsorship Decisions: Team Identification Matters.” Journal of Sport Management, vol. 30, no. 5, 2016, pp. 490–505. https://doi.org/10.1123/jsm.2015-0327.

“‘Chilling’: IOC President Blasts ‘Tremendous Coldness’ from Valieva’s Entourage after Skate.” NBC New York, 17 Feb. 2022, www.nbcnewyork.com/news/local/chilling-ioc-president-blasts-tremendous-coldness-from-valievas-entourage-after-skate/3560688/.

Coombs, W. Timothy, and Sherry J. Holladay. “Helping Crisis Managers Protect Reputational Assets: Initial Tests of the Situational Crisis Communication Theory.” Management Communication Quarterly, vol. 16, no. 2, 2002, pp. 165–186. https://doi.org/10.1177/089331802237233.

Court of Arbitration for Sport. “The Court of Arbitration for Sport Publishes the Arbitral Award Issued in the Proceedings Involving RUSADA, the ISU, WADA and Kamila Valieva.” 7 Feb. 2024.

Dedaj, Paulina. “Ex-Russian Olympic Coach Returns to Spotlight at 2026 Winter Games despite Doping Shadow.” Fox News, 6 Feb. 2026, www.foxnews.com/sports/ex-russian-olympic-coach-returns-spotlight-2026-winter-games-despite-doping-shadow.

Dunbar, Graham. “Russian Figure Skater Disqualified from 2022 Olympics in Doping Case Verdict.” PBS NewsHour, Associated Press, 29 Jan. 2024.

Figure Skating Federation of Russia. “Всероссийские соревнования ‘Гран При России Идель-2022’ [All-Russian Competition ‘Grand Prix of Russia Idel-2022’].” fsrussia.ru/sorevnovaniya/sorevnovaniya/vserossijskie-sorevnovaniya-gran-pri-rossii-idel-2022.

Figure Skating Federation of Russia. “Ростелеком - Чемпионат России 2023 [Rostelecom - Russian Championship 2023].” fsrussia.ru/sorevnovaniya/sorevnovaniya/rostelekom-chempionat-rossii-2023.

Figure Skating Federation of Russia. “Ростелеком - Чемпионат России 2024 [Rostelecom - Russian Championship 2024].” fsrussia.ru/sorevnovaniya/sorevnovaniya/rostelekom-chempionat-rossii-2024.

Figure Skating Federation of Russia. “Чемпионат России 2026 [Russian Championship 2026].” fsrussia.ru/sorevnovaniya/sorevnovaniya/chempionat-rossii-2026.

Grohmann, Karolos. “WADA Chief Uncomfortable with Former Valieva Coach at Milano Games.” Reuters, 5 Feb. 2026, www.reuters.com/sports/wada-chief-uncomfortable-with-former-valieva-coach-milano-g

ames-2026-02-05/.
International Skating Union. “ISU Decision Concerning the Participation of Limited Number of Individual Neutral Athletes (AIN) at Olympic Qualification Events under Strict Conditions.” 20 Dec. 2024, www.isu.org/news/isu-decision-concerning-the-participation-of-limited-number-of-individual-neutral-athletes-ain-at-olympic-qualification-events-under-strict-conditions/.
International Skating Union. “ISU Statement Olympic Winter Games Beijing 2022 Figure Skating Team Event – Final Ranking Confirmed.” 2 Aug. 2024, www.isu.org/news/isu-statement-olympic-winter-games-beijing-2022-figure-skating-team-event-final-ranking-confirmed/.
International Skating Union. “ISU Statement on the Ukrainian Crisis - Participation in International Competitions of Skaters and Officials from Russia and Belarus.” 1 Mar. 2022, www.isu.org/news/isu-statement-on-the-ukrainian-crisis-participation-in-international-competitions-of-skaters-and-officials-from-russia-and-belarus/.
International Skating Union. “Special Screening Process Completed for Individual Neutral Athletes (AIN) for the 2025/26 Olympic Qualification Events.” 13 May 2025, www.isu.org/news/special-screening-process-completed-for-individual-neutral-athletes-ain-for-the-202526-olympic-qualification-events/.
Mann, Brian. “Sports Tribunal Bans Russian Kamila Valieva from Figure Skating through 2025.” NPR, 29 Jan. 2024.
McLaren, Richard H. The Independent Person Report, Part II. World Anti-Doping Agency, 9 Dec. 2016.
Nivison, Austin. “Kamila Valieva Doping Scandal: RUSADA Appeals Decision Not to Ban Russian Figure Skater, Requests ‘Reprimand.’” CBS Sports, 24 Feb. 2023.
Sassenberg, Anne-Marie, and Cindy Sassenberg. “How Do Consumers Evaluate Sponsorship in the Aftermath of Sport Celebrity Scandals? Uncovering Evidence on Social Media Data Mining.” International Journal of Sports Marketing and Sponsorship, vol. 25, no. 5, 2024, pp. 1044–1058. https://doi.org/10.1108/IJSMS-09-2023-0200.
TASS. “ISU Clears Four Russian Figure Skaters for 2026 Olympics Qualifiers.” 13 May 2025, tass.com/sports/1957315.
“Who Is Adeliia Petrosian and Why Is She Such a Mystery in Figure Skating?” NBC New York, 16 Feb. 2026, www.nbcnewyork.com/olympics/2026-milan-cortina/who-is-adeliia-petrosian/6463316/.
Zaccardi, Nick. “ISU Names Figure Skaters from Russia Eligible for Olympic Qualifying as Neutral Athletes.” NBC Sports, 13 May 2025, www.nbcsports.com/olympics/news/russia-figure-skaters-2026-olympics-adelia-petrosian.

Data Quality as A Foundation for Reliable Artificial Intelligence By Zara Vivekanand

Abstract

Artificial intelligence systems increasingly influence decisions in healthcare, finance, education, transportation, employment, and information access. As these systems become more capable, much attention is directed toward model architecture, computational power, and algorithmic innovation. However, the reliability of an artificial intelligence system depends fundamentally on the quality of the data used to train, validate, test, and operate it. A sophisticated model cannot reliably compensate for data that are inaccurate, incomplete, unrepresentative, outdated, inconsistently labeled, duplicated, or poorly documented. Poor-quality data can reduce predictive accuracy, reproduce social biases, distort model evaluation, create hidden technical debt, and cause systems that initially perform well to deteriorate after deployment. This paper examines data quality as a foundation for reliable artificial intelligence by analyzing major dimensions of data quality and the mechanisms through which data problems affect machine-learning systems. It examines research on label errors, demographic representation, healthcare algorithms, data cascades, distribution shift, and large language models. The paper also considers methods for improving data quality through documentation, validation, representative sampling, annotation management, monitoring, and governance. It argues that organizations should treat data quality as a continuous lifecycle responsibility rather than a preliminary cleaning task. Reliable artificial intelligence requires not only advanced models but also data infrastructures capable of producing, documenting, evaluating, and maintaining information that is accurate and appropriate for the system's intended use.

Introduction

Recent advances in artificial intelligence have created a widespread impression that improved model architectures and increased computational scale are the primary drivers of AI performance. Modern machine-learning systems can classify images, recommend products, translate languages, predict clinical outcomes, and generate human-like text. Large language models have further strengthened the view that increasingly powerful models can extract useful patterns from enormous quantities of information. Yet every machine-learning model is ultimately dependent on the information from which it learns. Models do not observe the world directly. They observe datasets constructed through measurement, collection, labeling, filtering, and selection processes. If those processes produce distorted representations of reality, the model may learn those distortions as though they were legitimate patterns.

Consequently, AI reliability is inseparable from data quality. The National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework describes trustworthy AI in terms that include validity, reliability, safety, accountability, transparency, privacy, and fairness. These characteristics cannot be achieved solely through model design because each depends in part on the information entering the system throughout its lifecycle

(Tabassi). Similarly, European regulation has explicitly recognized the relationship between data quality and AI risk. Article 10 of the European Union's Artificial Intelligence Act requires training, validation, and testing datasets used for high-risk AI systems to be relevant, sufficiently representative, as complete and error-free as reasonably possible, and appropriate for the context in which the system will operate (European Parliament and Council). The provision demonstrates that data quality has moved beyond a technical concern and become part of the broader governance of artificial intelligence.

This emphasis reflects a fundamental property of machine learning: algorithms identify statistical patterns in examples. They generally cannot determine independently whether the examples themselves accurately represent the phenomenon being modeled. A mislabeled medical image can teach a model the wrong relationship between visual characteristics and disease. A hiring dataset created from historically unequal employment decisions can cause an algorithm to reproduce those inequalities. A dataset collected in one geographic or demographic environment may not represent another. Even an accurate dataset can become unreliable over time as populations, behavior, technologies, or economic conditions change.

Research on real-world AI development suggests that these problems are not unusual exceptions. Sambasivan et al., based on interviews with 53 AI practitioners working across several countries, described "data cascades" as compounding downstream problems caused by weaknesses in data practices. Such cascades appeared in 92 percent of the projects studied. Importantly, the researchers found that data work was frequently undervalued compared with model development even though data problems could affect every later stage of an AI system (Sambasivan et al.).

This paper argues that high-quality, fit-for-purpose data form a necessary foundation for reliable artificial intelligence because data quality determines what a model can learn, how accurately its performance can be measured, how fairly it behaves across populations, and whether its performance remains stable after deployment. However, data quality is not sufficient by itself. Model design, security, human oversight, evaluation, and governance also determine reliability. The importance of data quality therefore lies not in the claim that perfect data automatically produce perfect AI, but in the stronger observation that serious data defects create reliability problems that better algorithms alone cannot consistently solve.

Understanding Data Quality in Artificial Intelligence

Data quality is sometimes reduced to the concept of "clean data," but this description is too narrow for artificial intelligence. A dataset can be free from obvious formatting errors while still being unsuitable for a particular application. Data quality is better understood as fitness for purpose: whether information has the characteristics necessary to support the decisions and predictions for which it will be used.

Several dimensions are particularly important. Accuracy refers to whether observations and labels correctly describe reality. Completeness concerns whether important information is missing. Consistency measures whether data are recorded according to compatible definitions, units, and formats. Uniqueness concerns unnecessary duplicates. Timeliness refers to whether the

information remains current. Representativeness concerns whether relevant groups and conditions are sufficiently reflected in the dataset. Relevance asks whether the variables actually measure the phenomenon the system is intended to predict. Finally, provenance and documentation establish where the information came from, how it was collected, what transformations were applied, and what limitations accompany its use.

These dimensions often interact. Suppose an AI system is intended to predict whether patients are at risk for complications following surgery. A hospital might possess millions of accurate medical records, apparently providing an ideal dataset. However, if most records come from one hospital system, the dataset may not adequately represent patients treated elsewhere. If some demographic groups historically had less access to healthcare, the medical records may contain systematically different amounts of information across populations. If definitions of variables changed between hospitals, combining the data could create consistency problems. If an administrative variable such as healthcare cost is substituted for a clinical measure of illness, the dataset may contain accurate values but still encode the wrong target.

The last example is particularly important because data quality cannot be evaluated only by asking whether individual values are correct. A variable can be accurately measured yet conceptually inappropriate. The problem is therefore partly epistemological: developers must determine whether the data actually represent the concept that the model is supposed to learn.

For supervised machine learning, label quality is another major concern. Models are commonly trained on examples paired with target labels: an image might be labeled "malignant" or "benign," an email "spam" or "not spam," or a transaction "fraudulent" or "legitimate." Errors in these labels introduce noise directly into the learning process. Frénay and Verleysen found that label noise can decrease predictive performance, increase model complexity, and increase the amount of training data needed to learn successfully (Frénay and Verleysen).

The problem also affects evaluation. Northcutt, Athalye, and Mueller investigated test sets from ten widely used machine-learning datasets and estimated that at least 3.3 percent of their labels were erroneous on average. They estimated that at least 6 percent of the ImageNet validation set contained label errors. Because test sets are supposed to provide an independent measurement of performance, errors in those datasets can make one model appear superior to another for reasons unrelated to actual capability (Northcutt et al.).

This illustrates a broader principle: data quality matters during training, validation, testing, and deployment. Training data determine what patterns a model learns. Validation data affect model selection and parameter tuning. Test data determine the reported estimate of performance. Production data determine what the deployed system actually encounters. Reliability therefore cannot be secured by cleaning one training dataset at the beginning of a project.

How Poor Data Quality Undermines AI Reliability

Measurement Error and Incorrect Labels

One of the most direct pathways from poor data to unreliable AI is incorrect measurement. Machine-learning systems operate under the assumption that recorded variables contain useful information about the underlying phenomenon being modeled. If sensors are poorly calibrated, records are entered incorrectly, labels are inconsistent, or measurements differ across institutions, the model may learn patterns created by the data-generation process rather than by the phenomenon itself.

Label errors are particularly damaging because they create contradictory learning signals. If similar examples are assigned different labels because annotators disagree or make mistakes, the model must attempt to fit noise along with meaningful relationships. More complicated models may sometimes memorize these irregularities rather than generalizing from stable patterns.

The effects of label errors also challenge the assumption that greater model complexity always improves reliability. Northcutt et al. found that correcting mislabeled benchmark examples could alter comparisons between models. In some scenarios involving noisy data, smaller models became preferable to higher-capacity models after labels were corrected. Their research demonstrates that benchmark rankings may partially reflect properties of the dataset rather than genuine differences between algorithms (Northcutt et al.).

Missing Data and Selection Bias

Missing information introduces another problem. Missingness is not necessarily random. Individuals who interact more frequently with a healthcare system, online platform, or financial institution may generate more data than people who interact less frequently. As a result, a dataset can overrepresent individuals who are easiest to observe.

This distinction is critical because machine-learning models generalize most reliably when deployment conditions resemble training conditions. If the training sample excludes or inadequately captures important parts of the intended population, aggregate accuracy can conceal substantial errors for particular groups.

Imagine a speech-recognition system trained primarily on recordings from speakers with a narrow range of accents. The system could achieve high average accuracy on a similarly constructed test set but perform poorly for people with accents that are rare in the data. The mathematical algorithm might function exactly as designed. The reliability failure would originate from the population represented during development.

For this reason, data quantity should not be confused with data coverage. Ten million records drawn from a narrow population may be less useful for a particular deployment than one million carefully sampled records representing the full range of expected users and circumstances.

Representation and Bias

The relationship between data representation and model performance was demonstrated clearly in the Gender Shades study by Buolamwini and Gebru. The researchers examined

commercial gender-classification systems and identified major differences in accuracy across groups defined by gender and skin type. They also found that two commonly used facial-analysis datasets were overwhelmingly composed of lighter-skinned individuals: 79.6 percent in one dataset and 86.2 percent in another. In their evaluation, darker-skinned women experienced error rates as high as 34.7 percent, while the maximum error rate for lighter-skinned men was 0.8 percent (Buolamwini and Gebru).

The study is important because it demonstrates that a single measure of overall model accuracy can be misleading. Suppose a demographic group represents only a small fraction of the evaluation dataset. A model may perform poorly for that group while maintaining an impressive aggregate score. The system may therefore be described as highly accurate while remaining unreliable for a subset of its users.

Representation is not simply a requirement to make datasets demographically equal in every situation. The appropriate distribution depends on intended use. A disease occurring disproportionately in a particular population might legitimately require more examples from that population. The central principle is that sampling decisions should be purposeful, documented, and connected to the conditions in which the AI system will operate.

Poor Proxies and Historical Bias

Even representative datasets may contain misleading relationships. Machine-learning systems often use proxy variables because the true concept of interest is difficult to measure. The quality of the proxy can determine whether the resulting predictions are meaningful.

Obermeyer et al. provided a major example in healthcare. They examined an algorithm used to identify patients who might benefit from additional care management and found substantial racial bias. The system used healthcare costs as a proxy for healthcare needs. Because Black patients historically incurred lower healthcare spending than equally sick White patients, cost was not an unbiased measure of illness. Consequently, the algorithm underestimated the health needs of Black patients (Obermeyer et al.).

This case shows why data quality must include conceptual validity. The healthcare-cost data were not necessarily inaccurate. The problem was that cost did not represent health need equally across populations. An algorithm optimized using the variable could therefore accurately learn a statistical relationship while producing a socially and clinically misleading decision rule. AI developers consequently need domain expertise alongside technical expertise. Determining whether a variable is an acceptable proxy for illness, creditworthiness, educational potential, job performance, or another complex human characteristic cannot be solved through data cleaning alone. It requires understanding how the data were generated and which social processes influence them.

Data Quality as a System and Lifecycle Problem

Many AI projects initially treat data preparation as a stage that ends once a dataset is cleaned and a model is trained. Real-world systems challenge this assumption because data continually change.

Sculley et al. described machine-learning systems as particularly vulnerable to forms of "technical debt." Their analysis identified data dependencies, hidden feedback loops, unstable inputs, and changes in the external world as sources of long-term maintenance difficulty. Unlike conventional software, where behavior is determined primarily by programmed logic, an ML system's behavior is also shaped by its data environment. A model can therefore deteriorate even when its code has not changed (Sculley et al.).

Consider a fraud-detection system trained on historical transaction patterns. Criminal behavior changes in response to new security systems. Consumer payment methods change. New products appear. Economic conditions shift. A model that accurately distinguished fraudulent transactions in one year may encounter a meaningfully different population of transactions two years later. The original dataset did not necessarily contain poor-quality information when collected. It has become less relevant to current conditions.

This problem is commonly described using terms such as dataset shift, data drift, and concept drift. Dataset or covariate shift occurs when the distribution of inputs changes. Concept drift occurs when relationships between inputs and outcomes change. Research reviewing concept drift emphasizes that machine-learning performance can deteriorate when changing data distributions are not detected and addressed (Lu et al.). Monitoring and retraining are therefore important responses when production environments change.

This means timeliness is part of data quality. A dataset is not permanently "high quality" simply because it passed an audit at the time it was created. Organizations need processes for continuously determining whether production conditions remain consistent with assumptions made during development.

Sambasivan et al. 's concept of data cascades reinforces this lifecycle perspective. Their research suggests that small compromises made during collection or labeling can propagate throughout the system. A poorly defined label can influence training, cause unexpected model behavior, require later engineering workarounds, and eventually create failures that appear to be model problems. Because the effect appears downstream from its cause, organizations may spend considerable effort modifying algorithms when the underlying defect originates in the data (Sambasivan et al.).

Reliability therefore depends on organizational design as much as on technical procedure. Teams need clear responsibility for data sources, methods for reporting quality problems, and incentives to invest in data infrastructure. If organizations reward improvements in model accuracy but treat data documentation and maintenance as secondary work, they may systematically underinvest in the foundation on which the models depend.

Data Quality in Generative Artificial Intelligence

Generative artificial intelligence expands the data-quality problem because large language models and related systems are trained on datasets vastly larger and less controlled than those used in many traditional machine-learning applications. Instead of relying entirely on carefully labeled examples, foundation models may learn from large collections of web pages, books, code, images, and other materials.

Scale provides important benefits. Large and diverse corpora allow models to learn language, concepts, and relationships across many domains. However, increasing the quantity of data also increases the difficulty of understanding exactly what the dataset contains.

One issue is duplication. Lee et al. examined language-model training data and found extensive repetition and near-duplicate text. Their research showed that deduplication reduced memorization and improved the efficiency of training. They also found that overlap between training and validation data could distort evaluation. In their experiments, deduplication reduced the frequency with which models reproduced memorized training text and reduced train-test overlap, supporting the argument that careful data curation can improve both model behavior and confidence in evaluation (Lee et al.).

This result challenges the simplistic idea that more data are always better. If a dataset contains the same material thousands of times, repeated examples alter the statistical importance assigned to that content. The effective distribution of the dataset is different from what a raw record count suggests. Duplicate information can also increase the likelihood that a model memorizes material rather than learning generalizable patterns.

A related issue is benchmark contamination. Large language models may be trained on public internet data that contain questions or answers from benchmarks later used to evaluate those models. If evaluation examples appear in training data, strong benchmark performance may partly reflect memorization rather than generalization. Deng et al. investigated contamination in modern LLM benchmarks and noted concerns that benchmark results may overstate real-world capability when evaluation data overlap with pretraining corpora (Deng et al.).

Benchmark contamination is fundamentally a data-governance problem. Researchers need to know what information entered a model's training corpus and maintain separation between data used to teach the model and data used to evaluate it. That becomes difficult when training corpora contains billions or trillions of tokens collected from changing online sources.

Generative AI also illustrates an important limit to the thesis of this paper: better data do not eliminate every reliability problem. Large language models can generate statements that are unsupported or false even when their training data contain accurate information. Their objectives generally involve predicting plausible sequences rather than verifying every generated claim against an authoritative source. Consequently, hallucination and reasoning errors cannot be solved through data cleaning alone.

Nevertheless, data quality strongly affects the environment in which these problems occur. Outdated sources make current answers less reliable. Low-quality sources expose models to misinformation. Duplicate content may increase memorization. Missing domains limit

coverage. Benchmark leakage creates misleading evaluation. Poorly curated retrieval databases can cause retrieval-augmented systems to ground their answers in inaccurate documents. For these reasons, NIST's Generative Artificial Intelligence Profile extends its risk-management framework specifically to generative systems and emphasizes managing risks across the full AI lifecycle rather than treating model performance as an isolated engineering problem.

Generative AI therefore does not reduce the importance of data quality. It expands the meaning of data-quality engineering from cleaning structured tables to governing extremely large, heterogeneous, continuously changing information ecosystems.

Building High-Quality Data for Reliable AI

Improving data quality requires more than removing incomplete rows immediately before model training. It should begin with the definition of the AI problem itself.

Define the Intended Use

Organizations should first specify what the model is expected to predict, who will use it, which population it will affect, and under what conditions it will operate. These decisions establish the standards by which data quality should be evaluated.

A dataset may be appropriate for one task and inappropriate for another. Medical records from adults might support research on adult disease but not a pediatric decision system. Images collected during daylight might be sufficient for one computer-vision application and dangerous for a system expected to operate at night. Data quality is therefore inseparable from intended use.

This principle is reflected in the EU AI Act's requirements that datasets for high-risk systems be relevant to their intended purpose and account for the geographic, contextual, behavioral, or functional environment in which the system will operate (European Parliament and Council).

Establish Data Provenance

Developers should know where each important dataset originated. Provenance records can include the organization that collected the data, collection dates, sampling procedures, consent or licensing conditions, transformations, labeling methods, and known limitations. Gebru et al. proposed "datasheets for datasets" as a standardized approach to this problem. Drawing an analogy to datasheets used for physical components, they argued that datasets should be accompanied by documentation describing motivation, composition, collection processes, recommended uses, and other characteristics. Such documentation improves communication between dataset creators and users while encouraging accountability (Gebru et al.). Documentation is especially valuable when datasets are reused. A research dataset originally created for one purpose may later become training data for a completely different application. Without provenance information, future developers may not realize that the sampling process omitted important groups or that labels were created according to definitions that no longer apply.

Validate Data Automatically and Manually

Automated checks can identify many common problems: invalid values, duplicate records, missing fields, impossible dates, inconsistent units, unexpected categorical values, broken schemas, and sudden changes in distributions.

However, automated validation cannot determine every dimension of quality. A program can verify that an age variable contains numbers between zero and 120, but it cannot independently determine whether the dataset includes enough older adults for a particular healthcare application. It can detect that every image has a label, but it may not know whether the labels are clinically correct.

High-quality data programs therefore combine automated validation with human review and domain expertise.

Improve Labeling Procedures

For supervised-learning applications, organizations should establish explicit labeling guidelines. Ambiguous categories should be defined before large-scale annotation begins. Difficult cases can be labeled independently by multiple annotators and reviewed through adjudication. Measures of inter-annotator agreement can reveal categories for which the supposed "ground truth" is not actually consistent.

This is particularly important in domains involving subjective judgments. Labels such as "toxic," "high risk," "qualified," or "misleading" may require interpretation rather than simple observation. Disagreement among qualified annotators may indicate that the underlying task is ambiguous rather than that one annotator is wrong.

Reliable AI therefore sometimes requires acknowledging uncertainty in the labels themselves instead of forcing complex examples into overly simple categories.

Audit Representation and Performance by Subgroup

Dataset audits should compare the composition of training data with the population in which the system will operate. Developers should then evaluate performance separately for meaningful subgroups instead of relying exclusively on aggregate metrics.

The Gender Shades results demonstrate why this matters. A large performance disparity can remain hidden in an average score when one population constitutes only a small fraction of the dataset (Buolamwini and Gebru).

Subgroup evaluation should not be treated as a one-time fairness test. Performance should continue to be monitored after deployment, particularly when user populations change.

Prevent Leakage and Duplication

Training, validation, and test sets should be meaningfully independent. Exact duplicates and near-duplicates should be detected before evaluation. For generative AI, organizations may need additional techniques to determine whether public benchmarks have entered large pretraining corpora.

Lee et al.'s work demonstrates that deduplication can improve model training while producing more credible evaluation results (Lee et al.). As models increasingly learn from massive internet datasets, provenance and duplication detection should become standard elements of AI quality assurance rather than optional optimization steps.

Monitor Production Data

Deployment should initiate another stage of data-quality management rather than end it. Organizations should monitor changes in input distributions, missingness, error rates, model outputs, and subgroup performance.

Not every statistical shift requires retraining. A feature may change without affecting performance, while an important relationship can change even when simple input statistics appear stable. Monitoring should therefore connect data-quality measures to model outcomes whenever ground-truth feedback becomes available.

Models used for consequential decisions should also have procedures for escalation when data fall outside expected conditions. An AI system that cannot determine whether its inputs resemble the information on which it was validated may produce confident predictions in situations for which no reliable evidence exists.

Governance and Organizational Responsibility

Technical methods alone will not solve the data-quality problem if organizations do not assign responsibility for data.

Traditional software engineering often provides relatively clear ownership of code. Machine-learning systems introduce additional dependencies: data may be generated by separate departments, purchased from vendors, collected from users, derived from other algorithms, or transformed through complex pipelines. A failure in one upstream source can affect multiple downstream models.

Sculley et al. warned that data dependencies can create technical debt that is more difficult to identify than ordinary code dependencies (Sculley et al.). A reliable AI organization therefore needs data ownership, version control, lineage tracking, access rules, validation requirements, and mechanisms for communicating changes to downstream users.

Governance should also distinguish between data quality and data ethics while recognizing their overlap. A dataset could be technically accurate but obtained without appropriate consent. Another dataset could be legally collected yet systematically exclude a disadvantaged population. Reliability cannot be evaluated exclusively through predictive accuracy when the system affects people's opportunities, safety, or rights.

Regulatory developments increasingly reflect this broader understanding. The European Union's Artificial Intelligence Act places data-governance requirements on high-risk systems and specifically addresses representativeness, errors, completeness, bias detection, and relevant data gaps (European Parliament and Council). NIST's Artificial Intelligence Risk Management Framework similarly encourages organizations to manage AI risk across design, development,

deployment, and evaluation rather than treating trustworthiness as a final compliance test (Tabassi).

These frameworks suggest that organizations should treat data-quality evidence as part of demonstrating that an AI system is trustworthy. A developer should be able to explain not only which model architecture was used, but also why the training population was appropriate, how labels were verified, how test sets were separated, what known data limitations remain, and how changes will be monitored after deployment.

Limitations of a Data-Centered Approach

Emphasizing data quality can create its own oversimplification if it implies that every AI failure originates in data. Reliable AI is a system-level property.

An excellent dataset can still be used to train an inappropriate model. Developers can select the wrong objective function, evaluate the wrong metric, expose a system to adversarial manipulation, implement software incorrectly, or deploy a model in a context for which automated decision-making is unsuitable. Generative systems can produce unsupported statements because of their probabilistic generation mechanisms even when their datasets are well curated. Human users can also misuse correct predictions.

Furthermore, perfect data are generally impossible. Real-world measurements contain uncertainty. Medical experts disagree. Human behavior changes. Rare events are difficult to sample. Future conditions cannot be fully represented in historical datasets.

The objective should therefore not be to eliminate all imperfections before an AI system can be used. Instead, organizations should understand and manage uncertainty. Data limitations should be documented, consequential errors should be prioritized, and deployment decisions should reflect the severity of potential harm.

This distinction supports a more defensible version of the claim that data quality is foundational. A foundation does not determine every characteristic of the structure built upon it, but serious weaknesses in the foundation constrain the reliability of everything above it. Model architecture, evaluation, security, governance, and human oversight remain essential. Yet those mechanisms cannot consistently rescue an AI system whose underlying information systematically misrepresents the environment in which decisions will be made.

Conclusion

Artificial intelligence is often discussed as though intelligence resides primarily in algorithms. In practice, machine-learning systems derive their capabilities from interactions between algorithms, computational resources, human decisions, and data. Of these components, data are particularly fundamental because they provide the empirical representation of the world from which models learn.

Poor data quality can reduce predictive accuracy through measurement errors and incorrect labels. It can produce misleading benchmark results when evaluation datasets themselves contain mistakes. It can generate unequal performance when relevant populations are

underrepresented. It can reproduce historical inequalities when socially biased variables are treated as neutral indicators. It can make models deteriorate when changing real-world conditions make historical data obsolete. In large language models, duplication and benchmark contamination further demonstrate that simply increasing data volume does not guarantee better or more trustworthy systems.

Research across machine learning supports this conclusion. Frénay and Verleysen show the consequences of noisy labels. Northcutt et al. demonstrate that errors exist even in widely used benchmark datasets. Buolamwini and Gebru reveal major demographic disparities in facial-analysis performance. Obermeyer et al. show how an apparently reasonable proxy variable can encode structural inequality into healthcare decisions. Sambasivan et al. demonstrate how neglected data problems can cascade throughout AI development. Lee et al. show that data deduplication can improve language-model behavior and evaluation. Together, these findings indicate that many apparent "model problems" cannot be understood without examining the data-generating processes beneath them.

The appropriate response is not simply to clean datasets more aggressively. Organizations should approach data quality as a continuous discipline. They should define data requirements according to intended use, document provenance, evaluate representation, improve labeling practices, eliminate unnecessary duplication, separate evaluation data from training data, monitor changing production environments, and assign clear responsibility for data assets. Documentation methods such as datasheets for datasets can help make assumptions and limitations visible rather than allowing them to disappear inside complex pipelines.

At the same time, data quality should not be presented as a complete solution to AI reliability. Accurate and representative data can still be combined with poorly designed models or irresponsible deployment decisions. Reliable artificial intelligence requires high-quality data alongside appropriate algorithms, rigorous evaluation, transparency, security, governance, and human oversight.

Nevertheless, data quality occupies a distinctive position within this system. Models learn what their data make visible. They reproduce distinctions embedded in labels. They generalize from the populations represented during training. Their evaluation depends on the correctness of test datasets, and their continued reliability depends on whether future data remain sufficiently similar to past conditions. These constraints mean that advances in model architecture cannot eliminate the fundamental dependence of artificial intelligence on its informational foundation.

As AI systems become more deeply integrated into consequential decisions, organizations should therefore shift from a model-centered understanding of AI quality toward a systems-centered one in which data receive equal attention. The long-term reliability of artificial intelligence will depend not only on developing more powerful models, but on improving the processes through which machines are given information about the world. In that sense, data quality is not simply one technical consideration among many. It is a foundation upon which reliable artificial intelligence must be built.

Works Cited

Autio, Chloe, et al. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. National Institute of Standards and Technology, 2024. NIST AI 600-1, https://doi.org/10.6028/NIST.AI.600-1.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, vol. 81, 2018, pp. 77-91, proceedings.mlr.press/v81/buolamwini18a.html.

Deng, C., et al. "Investigating Data Contamination in Modern Benchmarks for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024, pp. 8706-19. https://doi.org/10.18653/v1/2024.naacl-long.482.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). 2024, eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Frénay, Benoît, and Michel Verleysen. "Classification in the Presence of Label Noise: A Survey." IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, 2014, pp. 845-69. https://doi.org/10.1109/TNNLS.2013.2292894.

Gebru, Timnit, et al. "Datasheets for Datasets." Communications of the ACM, vol. 64, no. 12, 2021, pp. 86-92.

Lee, K., et al. "Deduplicating Training Data Makes Language Models Better." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 8424-45. https://doi.org/10.18653/v1/2022.acl-long.577.

Lu, Jie, et al. "Learning under Concept Drift: A Review." IEEE Transactions on Knowledge and Data Engineering, 2020.

Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." Proceedings of the 35th Conference on Neural Information Processing Systems: Datasets and Benchmarks Track, 2021.

Obermeyer, Ziad, et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science, vol. 366, no. 6464, 2019, pp. 447-53. https://doi.org/10.1126/science.aax2342.

Sambasivan, Nithya, et al. "'Everyone Wants to Do the Model Work, Not the Data Work': Data Cascades in High-Stakes AI." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 2503-11.

Tabassi, Elham. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023. NIST AI 100-1, https://doi.org/10.6028/NIST.AI.100-1.

How Have FDA Breakthrough Therapy Designations for Psilocybin-assisted Therapy Between 2018–2024 Shifted Clinical Trial Design Standards and Psychiatrist Attitudes Toward Psychedelic Medicine in the US? By Srikar Kesanam

ABSTRACT

The US Food and Drug Administration granted Breakthrough Therapy designation to psilocybin for treatment-resistant depression in 2018 and for major depressive disorder in 2019. This narrative review examines how those designations affected clinical trial design standards and psychiatrist attitudes toward psychedelic medicine between 2018 and 2024. Findings are synthesized across three clinical domains and compared against conventional treatment benchmarks. The reviewed evidence shows a consistent clinical pattern across depression, anxiety, and addiction: rapid symptom onset within hours to days, durable response after one to three administrations, and adverse effects that are acute and session-bound rather than chronic. Effect sizes in the major depression trials substantially exceeded those typically reported for conventional pharmacotherapy, and abstinence outcomes in the addiction literature exceeded approved pharmacotherapy benchmarks by a wide margin. Trial design also changed measurably across the scope window. No Phase 2/3 or Phase 3 psilocybin trial was registered before 2022, and the post-designation period produced the field's first multi-site, actively controlled, dose-ranging trials. A 2023 national survey found 81 percent of responding psychiatrists agreed that classic hallucinogens show promise for psychiatric conditions, up from a 2016 baseline by the same research group. The designations appear to have accelerated a transition toward pharmaceutical-grade research infrastructure, though the underlying clinical signal predates them. Substantial barriers remain between existing trial results and accessible treatment, including unresolved functional unblinding, narrow and non-representative trial populations, absent facilitator training and credentialing systems, and reimbursement frameworks not built for a therapy of this structure. The evidence supports expanded access in supervised clinical settings for patients who have exhausted conventional options. It does not yet support broad prescribing.

KEYWORDS

Psilocybin, Breakthrough Therapy designation, treatment-resistant depression, psychedelic-assisted therapy, clinical trial design

INTRODUCTION

In 2018, the FDA granted Breakthrough Therapy designation to a Schedule I hallucinogen for the treatment of depression. It did so again the following year for a second indication. For a compound that had been federally classified since 1970 as having no accepted medical use and a high potential for abuse, this was a substantial reversal, and it happened with relatively little public attention.

The context for that decision is a mental health system under strain. Roughly one in five American adults experiences mental illness in a given year (National Alliance on Mental Illness,

n.d.), and depression is the leading cause of disability worldwide (World Health Organization, 2023). The pharmacological standard of care has not changed fundamentally in decades. SSRIs remain the first-line treatment for both depression and generalized anxiety, and they work reasonably well for a substantial portion of patients. For the rest, the picture is worse than most treatment guidelines admit. Between 30 and 50 percent of patients with major depressive disorder do not respond adequately to their first prescribed antidepressant, and 10 to 30 percent meet criteria for treatment resistance after two or more failed medication trials (Davis et al., 2021, JAMA Psychiatry). Those patients face a narrowing set of options, most of which require weeks of daily administration before anyone knows whether they are working. Addiction treatment is in a similar position. The approved pharmacotherapies for alcohol and tobacco dependence produce statistically real but modest effects, with adherence problems that erode even those.

Psilocybin is the psychoactive compound found in several species of mushrooms, used in indigenous religious and healing contexts for centuries and studied clinically in the United States through the 1950s and 1960s before the Controlled Substances Act effectively ended that research in 1970. It is a serotonergic psychedelic, meaning a psychoactive substance that induces an altered state of consciousness, and it functions primarily as an agonist on 5-HT2A receptors. An agonist is a substance that activates a cell receptor to induce a biological response. Those 5-HT2A receptors belong to the same receptor family that conventional antidepressants target through different mechanisms. Clinical research resumed in the 2000s at a small number of academic centers, largely funded by private foundations rather than government or industry sources.

The mechanism is easy to oversimplify, and doing so distorts what the clinical evidence can actually support. While 5-HT2A agonism is psilocybin's primary action, the pathway from receptor activation to clinical response is not linear. Reuptake inhibition, the standard mechanism behind SSRI treatment, is likewise only one step in a broader biochemical cascade, and even that works inconsistently across patients. Psilocybin's 5-HT2A agonism initiates a complex downstream process involving glutamatergic signaling, which is the primary communication process that triggers neuron firing, neuroinflammatory modulation, which is the regulation of immune response, and large-scale functional network reorganization. Clinical outcomes therefore reflect several interacting biological stages at once, which is part of why response varies as widely as it does between patients.

What changed in 2018 was not the evidence. Psilocybin worked, or did not, the same way before and after. What changed was the regulatory posture toward it. Breakthrough Therapy designation grants expedited FDA guidance, Fast Track eligibility, and a formal signal that a compound may offer substantial improvement over existing therapies. For a field that had been operating without any clear pathway to approval, that signal reorganized the incentives around psilocybin research almost immediately.

Whether psilocybin deserves serious attention is no longer much contested in the clinical literature. The active disagreement now is over whether the existing trial infrastructure is

rigorous enough to support the conclusions being drawn from it. That disagreement is what makes this question worth asking. If the designations improved trial quality, the case for psilocybin strengthened considerably. If they mainly accelerated volume without improving rigor, the field may be building toward an approval decision on evidence that cannot bear the weight.

This paper examines what followed. It reviews the clinical evidence for psilocybin across depression, anxiety, and addiction between 2018 and 2024, compares that evidence against conventional treatment benchmarks, and traces how trial design standards and psychiatrist attitudes shifted across the same period.

METHODOLOGY

Literature was identified through Google Scholar and PubMed using combinations of the terms psilocybin, psychedelic-assisted therapy, treatment-resistant depression, major depressive disorder, alcohol use disorder, tobacco dependence, Breakthrough Therapy designation, and clinical trial design. Inclusion was limited to peer-reviewed clinical trials, meta-analyses, registry analyses, and survey studies published in English. The scope window for the analytic argument is 2018 to 2024, with pre-2018 trials retained as background where they informed the regulatory decisions under examination. Screening and selection were carried out by a single reviewer.

LITERATURE REVIEW

Study	Indication	Design	N	Key outcome	Follow-up
Johnson et al. (2014)	Tobacco dependence	Open-label pilot	15	80% biologically verified abstinence	6 months
Bogenschutz et al. (2015)	Alcohol dependence	Open-label proof of concept	10	Abstinence rose only after first dosing session; effect sizes 0.75–1.38	36 weeks
Carhart-Harris et al. (2016)	Treatment-resistant depression	Open-label feasibility	12	8 of 12 in remission at 1 week; reductions sustained at 3 months	3 months
Griffiths et al. (2016)	Cancer-related anxiety and depression	Double-blind RCT, crossover, low-dose control	51	~80% with clinically meaningful reductions in anxiety and depressed mood	6 months
Ross et al. (2016)	Cancer-related anxiety and depression	Double-blind RCT, active control (niacin)	29	60–80% maintained clinically significant anxiety response	6.5 months
Johnson et al. (2017)	Tobacco dependence	Long-term follow-up of 2014 cohort	15	67% abstinent at 12 months; 60% at ~30 months	~30 months
Garcia-Romeu et al. (2019)	Problematic alcohol use	Retrospective online survey	343	83% no longer met severe AUD criteria after psychedelic use	Retrospective
Davis et al. (2021)	Major depressive disorder	Randomized, waiting-list controlled	27 randomized (24 analyzed)	71% response and 54% remission at 4 weeks; d = 2.5–2.6	4 weeks
Carhart-Harris et al. (2021)	Moderate-to-severe MDD	Phase 2 double-blind RCT, active comparator	59	No significant between-group difference on primary outcome; remission 57% vs. 28%	6 weeks
Bogenschutz et al. (2022)	Alcohol use disorder	Double-blind RCT, active placebo	93	Heavy drinking days 9.7% vs. 23.6%	32 weeks
Goodwin et al. (2022)	Treatment-resistant depression	Phase 2 double-blind dose-finding RCT	233	25 mg superior to 1 mg control; 10 mg not	3 weeks (primary), 12 weeks

a. Psilocybin and Depression

Psilocybin produces antidepressant effects through a cascade of neurological changes: increasing glutamate transmission, reducing brain inflammation, and most importantly, decreasing activity within the default mode network, a large-scale brain network that becomes highly active during wakeful, internally focused states of mind (Wang et al., 2024, Clinical Psychopharmacology and Neuroscience, 22(2), 222-231). The DMN is a set of interconnected brain regions associated with self-referential thought, rumination, and emotional regulation, all of which are chronically dysregulated in patients with major depressive disorder. Psilocybin's capacity to briefly suppress and then reorganize DMN activity is considered one of the primary drivers of its antidepressant effect, because it appears to disrupt entrenched negative thought patterns at the neurological level.

Before either designation, the evidence base was small but suggestive. Carhart-Harris et al. reported the first modern open-label feasibility study of psilocybin for treatment-resistant depression in 2016, in which twelve patients with moderate to severe TRD received two psilocybin doses seven days apart alongside psychological support. Depressive symptoms fell significantly at one week and remained reduced at three months, with eight of twelve patients meeting remission criteria one week after treatment (Carhart-Harris et al., 2016, Lancet Psychiatry). Those findings, alongside the cancer-anxiety trials discussed in section 5b, formed part of the preliminary clinical evidence supporting the FDA's decision to grant Breakthrough Therapy Designation to COMPASS Pathways for psilocybin in TRD in October 2018. An uncontrolled study of twelve patients is thin ground for a regulatory signal of that weight.

The post-designation record is substantially stronger. Davis et al. conducted a randomized, waiting-list-controlled trial of psilocybin-assisted therapy in adults with major depressive disorder, randomizing 27 participants to immediate or delayed treatment and analyzing the 24 who completed the intervention. Participants in the immediate treatment group showed significantly lower depression scores at both follow-up points than the delayed group did at equivalent timepoints, with Cohen's d values of 2.5 and 2.6. Seventy-one percent met criteria for clinically significant response at four weeks, and 54 percent were in remission (Davis et al., 2021, JAMA Psychiatry). The waiting-list design controls for the passage of time and for the effect of being accepted into a trial, but not for expectancy, which remains the central methodological weakness of the study.

The most significant clinical evidence to date came with the first head-to-head comparison between psilocybin and a gold-standard antidepressant, oral escitalopram. Carhart-Harris et al. conducted a phase 2 double-blind randomized controlled trial published in the New England Journal of Medicine, enrolling patients with long-standing moderate-to-severe major depressive disorder. Participants received either two doses of 25mg psilocybin three weeks apart plus six weeks of daily placebo, or two doses of 1mg psilocybin plus six weeks of daily oral escitalopram, a widely prescribed front-line SSRI. The primary outcome measure, change in depressive symptom severity at six weeks, did not differ significantly between groups, but secondary outcomes, including remission rates, emotional well-being, and ability to feel

pleasure, generally favored the psilocybin group (Carhart-Harris et al., 2021, NEJM). Remission at six weeks reached 57 percent in the psilocybin group compared to 28 percent for escitalopram. Psilocybin matched a daily six-week medication regimen using two administrations. That is the finding, more than any claim of superiority. Fewer doses and faster onset carry obvious clinical weight regardless of whether the two drugs separate statistically on a primary endpoint.

Building on this, a COMPASS Pathways-sponsored phase 2 double-blind dose-finding trial further refined the gray area surrounding psilocybin's clinical picture. Goodwin et al. found that a single 25mg dose of synthetic psilocybin reduced depression scores significantly more than a 1mg control dose at three weeks, while the 10mg dose did not separate from control (Goodwin et al., 2022, NEJM). That gave the field its first dose-ranging data in TRD, and the null result at 10mg matters as much for protocol design as the positive one at 25mg. More importantly, this trial also reinforced the rapid onset of psilocybin's antidepressant effects, a consistent feature across the literature that stands in sharp contrast to the four to six week delay usually associated with SSRI therapy.

The clinical evidence reviewed here supports psilocybin as a neurobiologically reasonable and clinically promising treatment for depression. Two NEJM publications and a JAMA Psychiatry trial in the space of eighteen months is not a specialty-journal literature anymore. The FDA's designations both reflected that shift and accelerated it.

b. Psilocybin's Impact on Anxiety

Psilocybin's neurobiological mechanism for anxiety relief operates largely through the same pathways already documented in the depression literature, though the specific clinical populations in which it has been most rigorously studied differ in ways that matter. As noted in the previous section, 5-HT2A receptor agonism decreases activity in the amygdala, the brain region primarily responsible for processing fear and threat responses, and promotes DMN stabilization. Dysregulation in both structures is not exclusive to MDD. It is also a consistent neurobiological feature of anxiety disorders (Wang et al., 2024, Clinical Psychopharmacology and Neuroscience). The mechanistic overlap is real, and it explains why anxiety became an adjacent research target so quickly. That said, this should not be read as evidence of a single unified pathway producing anxiety relief as a side effect of antidepressant action. Reduced amygdala reactivity likely involves distinct downstream processes from those driving DMN reorganization in depression, and the degree to which these separate processes contribute to observed anxiety-reducing outcomes is still not fully resolved in the literature.

The strongest clinical evidence for psilocybin's effects on anxiety comes from two double-blind, randomized controlled trials published simultaneously in the December 2016 issue of the Journal of Psychopharmacology, both focused on patients with cancer-related anxiety and depression. The first, conducted by Griffiths et al. at Johns Hopkins, enrolled 51 patients with life-threatening cancer diagnoses and assigned them in a crossover design to either a high therapeutic dose of psilocybin (22 or 30 mg/70 kg) or a very low placebo-equivalent dose (1 or 3 mg/70 kg), with sessions five weeks apart and follow-up at six months. At the six-month mark,

approximately 80 percent of participants showed clinically meaningful reductions in both depressed mood and anxiety, alongside self-reported improvements in quality of life, sense of meaning, and decreased fear of death (Griffiths et al., 2016, Journal of Psychopharmacology).

The second trial, led by Ross et al. at NYU, enrolled 29 cancer patients with anxiety and depression and compared a single moderate dose of psilocybin (0.3 mg/kg) against niacin as an active control, both administered alongside structured psychotherapy sessions. Psilocybin produced reductions in anxiety that were immediate in onset and sustained over time, with 60 to 80 percent of participants maintaining clinically significant anxiety-reducing responses at the 6.5-month follow-up (Ross et al., 2016, Journal of Psychopharmacology).

Both trials share a limitation worth stating plainly. Griffiths et al. and Ross et al. studied anxiety in patients with life-threatening cancer diagnoses, which is a specific kind of anxiety. Existential distress in someone facing a terminal prognosis is not obviously the same clinical target as generalized anxiety disorder, panic disorder, or social anxiety, and the mechanisms driving relief in one population may not transfer cleanly to the others. There is also a timing problem. Both trials were published in 2016, before either Breakthrough Therapy designation, and the anxiety literature has not seen anything like the post-designation expansion that followed in depression research. Norring and Spigarelli's trial registry analysis found 39 of 134 psilocybin trials targeting depression indications, while anxiety appears mostly as a secondary outcome measure rather than a primary endpoint (Norring and Spigarelli, 2024, DDDT). Within the population it was collected from, the anxiety evidence holds up well. Extension beyond that population has barely been attempted.

c. Psilocybin's Impact on Addiction

Psilocybin has been studied far less extensively in addiction than in depression or anxiety, and the mechanism at work looks meaningfully different from the antidepressant one. Two separate pathways appear to drive abstinence outcomes, and they operate differently from one another.

The first is biological. 5-HT2A agonism and the resulting disruption of DMN activity are again implicated, which produces the same downstream reorganization of self-referential thought documented in the depression literature (Wang et al., 2024, Clinical Psychopharmacology and Neuroscience). But addiction introduces a variable the depression and anxiety literature does not have to account for. Craving and compulsive use are driven substantially by dopaminergic reward circuitry rather than by the serotonergic and glutamatergic pathways discussed above. Psilocybin's effects on reward learning and habit-based behavior remain comparatively understudied at the mechanistic level, which means the biological account of how psilocybin reduces craving is incomplete even where the clinical outcomes are consistent.

The second is psychological. Across multiple trials, the intensity of the mystical-type or spiritually significant experience during dosing sessions predicts the durability of abstinence outcomes. That is a distinct causal pathway with no analogue in conventional pharmacotherapy, where the subjective experience of taking a drug is generally irrelevant to whether it works.

Psilocybin contributed to a measurable improvement in abstinence among patients with alcohol use disorder, and two trials from the same research group support that conclusion through different methods. Bogenschutz et al. first tested the compound in an open-label proof-of-concept study at the University of New Mexico, where ten participants with DSM-IV alcohol dependence received one or two psilocybin sessions (0.3 to 0.4 mg/kg) alongside motivational enhancement therapy. Abstinence did not improve during the first four weeks, when participants received counseling but no psilocybin, then increased significantly after the first psilocybin session and was largely sustained through 36 weeks of follow-up, with effect sizes ranging from 0.75 to 1.38 (Bogenschutz et al., 2015, Journal of Psychopharmacology). The design cannot establish efficacy on its own. Ten participants and no control group make this a feasibility signal, but the timing of the improvement, arriving only after dosing rather than after counseling alone, is difficult to attribute to the therapy component.

The same group later tested that signal under controlled conditions. Their double-blind RCT at two academic centers compared psilocybin against diphenhydramine as an active placebo rather than an inert one, a choice that addresses a problem specific to psychedelic trials, which is that participants can usually identify which arm they are in from the drug's subjective effects alone. Ninety-three patients with AUD received psychotherapy plus either psilocybin (25 to 40 mg/70 kg across two sessions) or diphenhydramine. Over 32 weeks, the psilocybin group's percentage of heavy drinking days was 9.7 percent against 23.6 percent for diphenhydramine, a gap of nearly 14 percentage points (Bogenschutz et al., 2022, JAMA Psychiatry). Mean daily alcohol consumption was lower in the psilocybin group as well, and no serious adverse events occurred among those who received it. Taken together, the open-label and controlled findings converge on the same causal direction despite the acknowledged weaknesses of the earlier design.

The evidence isn't limited to alcohol. Johnson et al. at Johns Hopkins ran an open-label pilot in 2014 testing psilocybin as an addition to a 15-week smoking cessation program. Fifteen nicotine-dependent smokers, with a mean of six prior failed quit attempts and 31 years of smoking on average, received two to three psilocybin sessions (20 to 30 mg/70 kg) alongside cognitive behavioral therapy. Twelve of fifteen, or 80 percent, showed biologically verified abstinence at six months (Johnson et al., 2014, Journal of Psychopharmacology). A follow-up of the same cohort found 67 percent still abstinent at 12 months and 60 percent abstinent at an average of 30 months post-treatment (Johnson et al., 2017, American Journal of Drug and Alcohol Abuse). No control group, a small sample, and self-selected participants mean the abstinence rate cannot be pinned on psilocybin with the confidence a randomized trial would give. The effect did hold across three separate follow-up windows.

Naturalistic survey data speaks directly to the second mechanism. Garcia-Romeu et al. surveyed 343 adults who reported reducing or stopping problematic alcohol use after psychedelic use outside a clinical setting, mostly LSD or psilocybin. Seventy-two percent met retrospective criteria for severe AUD before their psychedelic experience, and 83 percent no longer met those

criteria afterward. Higher psychedelic dose, more intense mystical-type experiences, and greater personal meaning attributed to the experience all tracked with larger reductions in drinking, even controlling for how severe the drinking was beforehand (Garcia-Romeu et al., 2019, Journal of Psychopharmacology). Survey data cannot establish causality, and people willing to complete a survey about a positive psychedelic experience are not a random sample. Still, the dose-response pattern and the association with mystical experience align with the controlled trial findings closely enough to corroborate the psychological pathway, even where they cannot independently establish it.

Whether the biological mechanism alone would produce these outcomes without the psychological one is not something the current literature can answer. The two are confounded in every trial reviewed here, because every trial administered psilocybin in a supported therapeutic setting designed to facilitate exactly the kind of experience that predicts better outcomes.

d. Comparison with Traditional Treatment Methods

Decades of SSRI dominance in first-line treatment have produced a system that is safe and widely accessible but leaves a large share of patients without adequate relief. That gap in coverage, not the dominance itself, is the reason psilocybin's higher response rates deserve close reading. A meta-analysis published in The Lancet found that while all 21 antidepressants studied outperformed placebo, effect sizes were small to moderate across the board (Cipriani et al., 2018, The Lancet). In practice, roughly 30 to 50 percent of patients with MDD do not respond to their first prescribed antidepressant, and somewhere between 10 and 30 percent are considered treatment-resistant after failing two or more adequate medication trials (Davis et al., 2021, JAMA Psychiatry). For those patients, the options expand but do not improve dramatically: electroconvulsive therapy, transcranial magnetic stimulation, ketamine infusions, or switching drug classes, each with its own adherence problems, access barriers, and inconsistent durability.

There is also a time problem that gets normalized in clinical practice. SSRIs typically require four to six weeks of daily administration before therapeutic benefit appears, during which patients are still symptomatic and often managing side effects on top of the condition the drug is meant to treat. A treatment that requires daily medication for six weeks just to determine whether it is working demands a level of sustained motivation and stability that many people in a depressive episode do not reliably have.

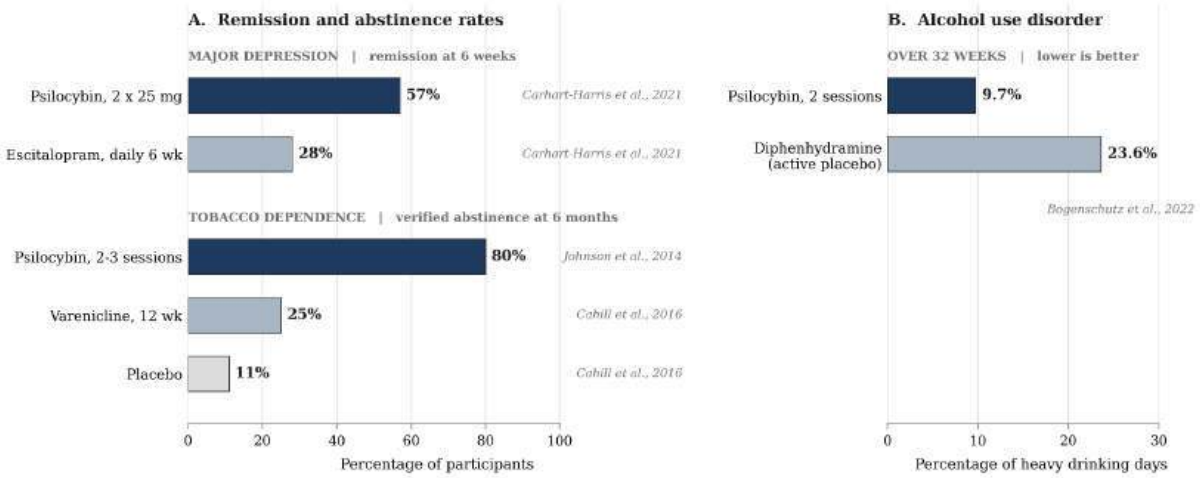
Psilocybin compresses that timeline from weeks to a single day and reduces total administrations from roughly forty to fewer than three. In the Davis et al. randomized waitlist-controlled trial, the antidepressant effect was detectable by day one after the first session, with QIDS-SR depression scores dropping from a mean of 16.7 at baseline to 6.3 the following day, a Cohen's d of 2.6 that held through the four-week follow-up (Davis et al., 2021, JAMA Psychiatry). In the Carhart-Harris escitalopram trial, psilocybin's QIDS score change from the day before to the day after the first dosing session was one of the few between-group differences that reached statistical significance, meaning psilocybin's onset advantage is faster not just than SSRIs in general but faster than escitalopram within the very first treatment window

(Carhart-Harris et al., 2021, NEJM). Whether that speed advantage carries clinical meaning over longer follow-up periods is still unclear, but the difference in treatment burden is not ambiguous.

The adverse effect comparison generates persistent misunderstanding, and the cause is that psilocybin and SSRIs distribute their side effects differently across time. Psilocybin concentrates its adverse effects into the dosing session itself: transient blood pressure and heart rate increases, headache, nausea, and periods of psychological distress that sometimes require active facilitator support. Because those effects are acute and visible, they register as more alarming than the diffuse, cumulative burden of daily medication. SSRIs spread their side effects across months, which makes each individual instance easier to dismiss even as the total burden accumulates. In the Carhart-Harris escitalopram trial, four of 29 escitalopram patients stopped taking their daily capsules entirely because of side effects, while no psilocybin patient requested cancellation of their second dose (Carhart-Harris et al., 2021, NEJM). Overall adverse event rates were similar across both arms, at 87 percent for psilocybin and 83 percent for escitalopram. The distribution differed, not the frequency.

For anxiety, the standard comparators are SSRIs and SNRIs for generalized presentations, with benzodiazepines covering acute management despite the dependency risk that makes long-term use impractical. Neither class produces sustained responses at the magnitude seen in the Griffiths and Ross 2016 cancer trials, where roughly 60 to 80 percent of participants maintained clinically meaningful anxiety reductions six months or more after one or two sessions, with no dependency profile (Griffiths et al., 2016; Ross et al., 2016, Journal of Psychopharmacology).

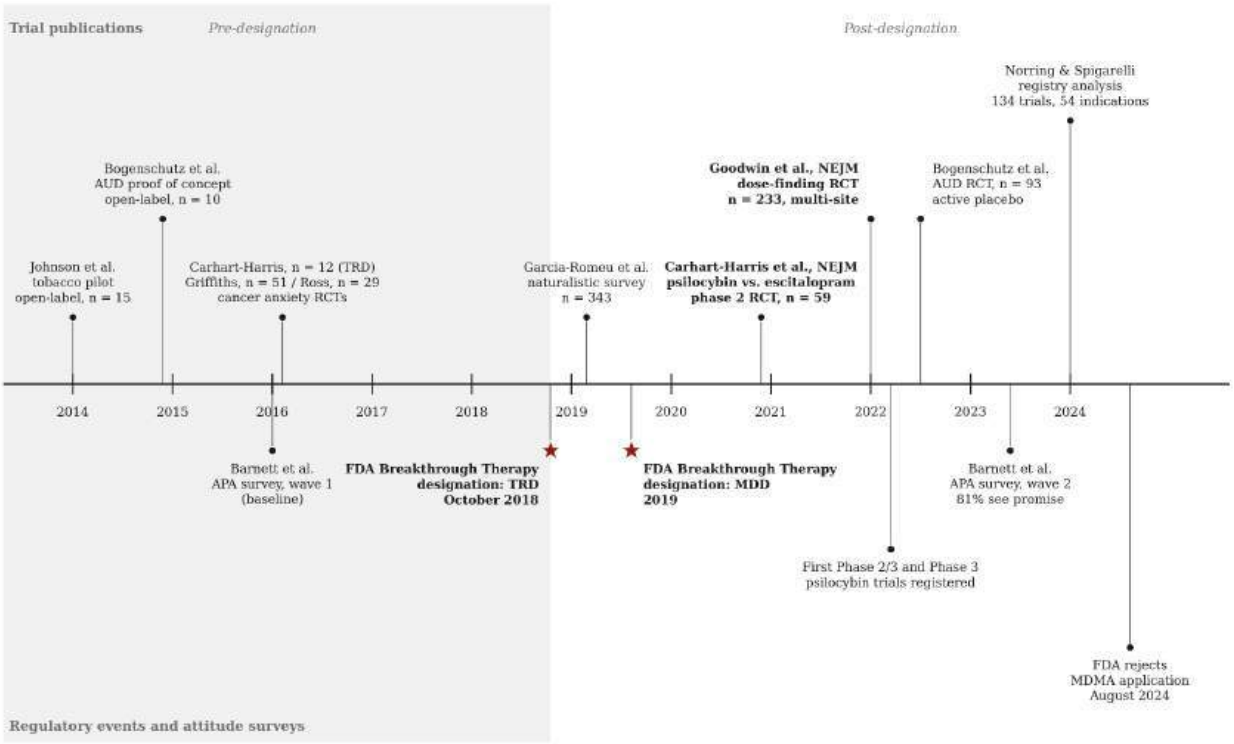
For addiction the contrast is starker, because the approved pharmacotherapies are not very effective in absolute terms. A meta-analysis of 64 randomized controlled trials found that naltrexone and acamprosate, the two most widely prescribed medications for alcohol use disorder, produced a combined effect size of Hedges' g = 0.209 against placebo. In practical terms, nine people need to be treated with naltrexone to prevent one additional return to heavy drinking, and eight with acamprosate to achieve one additional case of abstinence (Maisel et al., 2013, Addiction). For smoking, the best available pharmacotherapy, varenicline, showed sustained abstinence at six months in approximately 25 percent of treated participants versus 11 percent on placebo across 27 randomized controlled trials (Cahill et al., 2016, Cochrane Database of Systematic Reviews). Against those benchmarks, the 80 percent six-month abstinence rate in Johnson et al.'s tobacco pilot and the 14 percentage-point gap in heavy drinking days in Bogenschutz et al.'s 2022 AUD RCT are a different order of magnitude entirely, even accounting for the limitations of open-label designs and tightly controlled trial populations.



The obvious limitation across all these comparisons is that psilocybin has only been tested head-to-head against an active comparator once, in the Carhart-Harris depression-escitalopram trial. Without direct comparisons for anxiety and addiction, the apparent effect size differences between psilocybin trials and conventional treatment trials have to be read carefully, because different patient populations, outcome measures, and follow-up windows can easily inflate what looks like a dramatic gap. What can be said is that psilocybin's treatment model is structurally distinct from conventional pharmacotherapy in ways patients directly experience: fewer administrations, faster onset, a time-limited rather than chronic side effect burden, and no serious adverse events attributed to psilocybin across any of the major trials reviewed here.

DISCUSSION

Within four years of the 2018 designation, two psilocybin trials had been published in the New England Journal of Medicine. The Goodwin et al. 2022 COMPASS trial alone enrolled 233 participants across multiple sites with a proprietary synthetic formulation and a dose-ranging control design, a level of trial sophistication that reflects what happens when a development program has real regulatory scaffolding behind it rather than philanthropic funding alone (Goodwin et al., 2022, NEJM). Norring and Spigarelli's 2024 ClinicalTrials.gov analysis confirmed the timing. Before 2022, not a single Phase 2/3 or Phase 3 psilocybin trial had been submitted, and all three advanced trials appeared in the last two years of their study window (Norring and Spigarelli, 2024, DDDT). Before the designations, the picture looked very different. The Griffiths et al. and Ross et al. cancer anxiety trials published in 2016 were serious work, but they appeared in a specialty psychopharmacology journal, enrolled narrow patient populations, and operated without any regulatory pathway to approval. The designations did not cause psilocybin to work. They caused the field to test it the way it needed to be tested.



Psychiatrist attitudes shifted alongside the research, and there is survey data on the shift rather than just impression. Barnett and colleagues surveyed randomly selected members of the American Psychiatric Association in 2016, then repeated the survey in 2022 and 2023 using virtually identical methods. Because the first survey predates the 2018 designation and the second follows it, the pair functions as a rough before-and-after measure. The follow-up found 81 percent of responding psychiatrists agreed that psychedelics show promise in treating psychiatric conditions, with over half reporting they planned to incorporate psychedelic-assisted therapy into their practices upon FDA approval, a marked increase in optimism and a corresponding decrease in expressed concern about risk relative to the 2016 baseline (Barnett et al., 2018, Journal of Nervous and Mental Disease; Barnett et al., 2023, Psychedelic Medicine). The trials reviewed in sections 5a through 5d sit at the better-designed end of a field where much of the surrounding work does not meet that standard.

The clinical picture from those better-designed trials is difficult to dismiss. Across depression, anxiety, and addiction, the pattern is consistent: onset within hours to days rather than weeks, meaningful response after one to three sessions, and side effects concentrated in the dosing session instead of accumulating over months of daily use. The effect sizes in Davis et al. and Carhart-Harris et al. run substantially larger than what pharmacological depression trials typically report, and Johnson et al.'s 80 percent six-month tobacco abstinence rate has no real comparison point in the approved cessation literature (Davis et al., 2021, JAMA Psychiatry; Carhart-Harris et al., 2021, NEJM; Johnson et al., 2014, Journal of Psychopharmacology). For patients who have already cycled through conventional treatment without adequate response, the

risk-benefit calculation looks different than it does for a first-line prescribing decision. The evidence reviewed here makes a reasonable case that those patients, treated in supervised clinical settings with trained facilitators, are not well served by waiting another decade for perfect trial data before access expands.

The addiction evidence deserves separate weighting. One RCT, two open-label pilots, and a survey study is a thinner record than the depression or anxiety literature, where two NEJM papers landed within five years of each other. But it follows roughly the same trajectory psilocybin's depression research followed about a decade earlier, just starting later. Whether addiction receives the kind of FDA attention TRD and MDD have already received remains an open question, and the answer will likely determine whether the AUD findings get the confirmatory trials they need.

Across the trials reviewed here, psilocybin administered in controlled settings with appropriate screening carried a manageable safety profile and produced clinically meaningful effects in the populations studied. Whether those results survive the loosening of those conditions is a separate question, and the trial record has almost nothing to say about it. A well-run Phase 2 trial and a treatment that scales to a general psychiatric population are separated by a distance the existing evidence does not cross. The designations accelerated a real and important shift in how seriously this compound is taken, and the research that followed justifies that shift.

LIMITATIONS

Sources were identified and selected through a search-and-screening process modeled on PRISMA guidance, but this is a narrative review in its synthesis rather than a systematic review or meta-analysis, and no PRISMA flow diagram or record count is reported here. No effect sizes were pooled, no formal risk-of-bias instrument was applied to the included trials, and no quantitative synthesis was performed. A meta-analysis of the same source pool would be better positioned to detect patterns across the literature and to weight findings by methodological quality rather than by relevance to the argument being made.

Searches were run through Google Scholar and PubMed. Both are large and appropriate for clinical literature, but neither covers Embase, PsycINFO, or the Cochrane Library, all of which index psychiatric and psychopharmacological research that PubMed indexes incompletely or not at all. No grey literature search was conducted, meaning conference abstracts, dissertations, and unpublished trial data fall outside this review entirely. Screening and selection were also carried out by a single reviewer, without the dual independent screening or inter-rater reliability checks that reduce selection bias in systematic reviews. Where a judgment call determined whether a study was included, one person made it.

The final source base is small. The clinical claims in sections 5a through 5d rest on roughly a dozen primary trials, several with sample sizes under 30. Davis et al. analyzed 24 completed participants. Johnson et al.'s smoking cessation pilot had 15. Bogenschutz et al.'s 2015 proof-of-concept study had 10. These are real studies producing real data, but effect sizes

calculated on samples that small are unstable, and a paper building an argument on them inherits that instability. The larger trials, Goodwin et al. at 233 participants and Bogenschutz et al. 2022 at 93, carry more weight, and the argument here leans on them more heavily than the citation count suggests.

The 2018 to 2024 scope boundary cuts both ways. It keeps the analysis focused on the post-designation period, which is the point of the research question, but it excludes 2025 publications that would sharpen the picture, including newer alcohol use disorder RCTs. It also relegates foundational pre-2018 work to background rather than evidence. Any conclusion here is provisional against a literature still actively expanding.

The psychiatrist attitudes portion of the research question rests on thinner evidence than the clinical portion. The before-and-after comparison drawn here depends on two surveys by a single research group, and while the repeat survey used methods closely matched to the original, the response rates diverge sharply. The 2016 survey drew responses from 32 percent of the 1,000 psychiatrists contacted. The follow-up, working from a sample of the same size, drew 131. At that rate, respondents with strong existing opinions are likely overrepresented in either direction. No primary data was collected for this paper and no psychiatrists were interviewed, so the claim that professional attitudes shifted rests entirely on secondary reporting of those two instruments. The direction of the change is consistent with the surrounding evidence on funding, publication, and trial registration, but it is not established with the confidence the clinical findings are.

There is also a causal problem this paper cannot resolve. The correlation between the Breakthrough Therapy designations and the subsequent improvement in trial design and scale is clear in the timeline. Establishing that the designations caused that improvement is harder. Psychedelic research was already gaining momentum through the 2010s, private foundation funding was increasing independently, and cultural attitudes were shifting across the same period. The designations are the most visible inflection point, and the registry data showing zero Phase 2/3 or Phase 3 trials before 2022 is suggestive, but separating regulatory effect from broader trend would require an analysis this paper does not perform.

Finally, this review inherits the limitations of the trials it reviews. Functional unblinding remains unresolved across nearly all psychedelic research, including the studies treated here as the strongest evidence. Published trials skew positive by nature, and no method used here accounts for negative results that were never published. Every conclusion drawn is bounded by those constraints.

FURTHER RESEARCH

Several gaps in the existing literature are large enough to shape what the field should prioritize next.

The first is trial rigor, and specifically the underuse of randomized controlled designs. Norring and Spigarelli found that 50 percent of psilocybin trials initiated in the three years preceding their analysis were designed without blinding, a problem that inflates effect estimates and undermines the reliability of patient-reported outcomes (Norring and Spigarelli, 2024,

DDDT). Unblinded clinicians may unknowingly treat participants differently across arms, and unblinded participants may report improvement they expect rather than improvement they experience. Given that functional unblinding is already the central methodological criticism facing psychedelic research, and given that it featured directly in the FDA's rejection of the MDMA application, designing trials without blinding is a choice the field can no longer afford. Future research should also test alternatives to inert placebo, following the active-comparator approach Bogenschutz et al. used in 2022, since an inert placebo provides almost no blinding against a compound with effects this discriminable.

The second is population representativeness. The trial populations reviewed here were predominantly white, highly educated, medically screened, and treated in research settings with intensive support infrastructure. The patients who carry the heaviest burden of treatment-resistant depression, alcohol use disorder, and addiction in the United States are largely not the patients who were studied. Trials enrolling more racially and socioeconomically diverse participants, and trials that relax the medical exclusion criteria to include patients with the comorbidities common in real clinical practice, would do more to establish whether these results generalize than additional trials in the populations already well characterized. Until that work exists, the external validity of every finding in this review remains an open question.

The third is the research agenda itself. Norring and Spigarelli found 54 distinct indications across 134 registered trials. That is a field spreading itself thin. Concentrating funding and regulatory attention on a smaller number of well-powered, properly blinded trials in the indications with the strongest existing signal, which are treatment-resistant depression, major depressive disorder, and alcohol use disorder, would move psilocybin toward approval faster than continuing to test it against everything at once.

CONCLUSION

The FDA's Breakthrough Therapy designations did what they were designed to do. They pulled psilocybin research out of the margins and into a regulatory framework that demanded better trials, and the field responded with better trials. Building the system that would actually deliver this treatment to patients was always going to be someone else's job. What comes next depends on whether the infrastructure, the regulatory framework, and the clinical training can catch up to what the trial data already suggests is possible.

That gap is where policy attention should go now. Two things matter most.

The first is infrastructure. Psilocybin-assisted therapy requires trained facilitators for multi-hour supervised sessions, plus preparatory and integration meetings around them. No training pipeline currently exists at the scale approval would require, and no credentialing standard governs who is qualified to sit in that room. Oregon became the first state to permit supervised psilocybin therapy outside a research setting, and other states have since pursued their own legislative approaches (Wang et al., 2024, Clinical Psychopharmacology and Neuroscience). These programs will generate useful data, but they operate outside the FDA framework and their

standards vary. Federal guidance on facilitator qualifications should not wait until after approval to be written.

The second is reimbursement. A treatment that costs several thousand dollars per course and is not covered by insurance is not a treatment most people can access, whatever the trial data shows. The patients with the strongest case for psilocybin are disproportionately those who have already failed multiple conventional treatments, which correlates with longer illness duration, more disability, and less financial capacity. Building a delivery model that only works for people who can pay out of pocket would invert the logic that justified the designations in the first place.

This is not an argument for broad prescribing. Psilocybin-assisted therapy as currently designed requires multi-hour supervised sessions, preparatory and integration meetings, and clinical infrastructure that existing insurance reimbursement frameworks are not built to handle. The MDMA situation is a useful reminder of how quickly a promising psychedelic trial record can run into regulatory reality. The FDA rejected Lykos Therapeutics' New Drug Application in August 2024 despite two positive Phase 3 trials, citing blinding problems, incomplete safety data collection, and clinical oversight failures that are not unique to MDMA (Barnett et al., 2025, Cleveland Clinic Journal of Medicine). Psilocybin will face versions of the same scrutiny.

None of this resolves the harder question of whether psilocybin will work outside the conditions that produced these results. Trial participants were screened, supported, and treated in settings built for research. Real clinical practice will look different, and some of the effect will likely be lost in translation. The evidence reviewed here justifies finding out

Works Cited

Barnett, B. S., Arakelian, M., Beebe, D., Ontko, J., Riegal, C., Siu, W. O., Weleff, J., & Pope, H. G. (2023). American psychiatrists' opinions about classic hallucinogens and their potential therapeutic applications: A 7-year follow-up survey. Psychedelic Medicine, 2(1), 1–9. https://doi.org/10.1089/psymed.2023.0036

Barnett, B. S., Mauney, E. E., & King, F., IV. (2025). Psychedelic-assisted therapy: An overview for the internist. Cleveland Clinic Journal of Medicine, 92(3), 171–180. https://doi.org/10.3949/ccjm.92a.24032

Barnett, B. S., Siu, W. O., & Pope, H. G., Jr. (2018). A survey of American psychiatrists' attitudes toward classic hallucinogens. Journal of Nervous and Mental Disease, 206(6), 476–480. https://doi.org/10.1097/NMD.0000000000000828

Bogenschutz, M. P., Forcehimes, A. A., Pommy, J. A., Wilcox, C. E., Barbosa, P. C., & Strassman, R. J. (2015). Psilocybin-assisted treatment for alcohol dependence: A proof-of-concept study. Journal of Psychopharmacology, 29(3), 289–299. https://doi.org/10.1177/0269881114565144

Bogenschutz, M. P., Ross, S., Bhatt, S., Baron, T., Forcehimes, A. A., Laska, E., Mennenga, S. E., O'Donnell, K., Owens, L. T., Podrebarac, S., Rotrosen, J., Tonigan, J. S., & Worth, L. (2022). Percentage of heavy drinking days following psilocybin-assisted psychotherapy vs placebo in the treatment of adult patients with alcohol use disorder: A randomized clinical trial. JAMA Psychiatry, 79(10), 953–962. https://doi.org/10.1001/jamapsychiatry.2022.2096

Cahill, K., Lindson-Hawley, N., Thomas, K. H., Fanshawe, T. R., & Lancaster, T. (2016). Nicotine receptor partial agonists for smoking cessation. Cochrane Database of Systematic Reviews, 2016(5), Article CD006103. https://doi.org/10.1002/14651858.CD006103.pub7

Carhart-Harris, R. L., Bolstridge, M., Rucker, J., Day, C. M. J., Erritzoe, D., Kaelen, M., Bloomfield, M., Rickard, J. A., Forbes, B., Feilding, A., Taylor, D., Pilling, S., Curran, V. H., & Nutt, D. J. (2016). Psilocybin with psychological support for treatment-resistant depression: An open-label feasibility study. The Lancet Psychiatry, 3(7), 619–627. https://doi.org/10.1016/S2215-0366(16)30065-7

Carhart-Harris, R. L., Giribaldi, B., Watts, R., Baker-Jones, M., Murphy-Beiner, A., Murphy, R., Martell, J., Blemings, A., Erritzoe, D., & Nutt, D. J. (2021). Trial of psilocybin versus escitalopram for depression. New England Journal of Medicine, 384(15), 1402–1411. https://doi.org/10.1056/NEJMoa2032994

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., Ruhe, H. G., Turner, E. H., Higgins, J. P. T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J. P. A., & Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. The Lancet, 391(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

Davis, A. K., Barrett, F. S., May, D. G., Cosimano, M. P., Sepeda, N. D., Johnson, M. W., Finan,

P. H., & Griffiths, R. R. (2021). Effects of psilocybin-assisted therapy on major depressive disorder: A randomized clinical trial. JAMA Psychiatry, 78(5), 481–489. https://doi.org/10.1001/jamapsychiatry.2020.3285

Garcia-Romeu, A., Davis, A. K., Erowid, F., Erowid, E., Griffiths, R. R., & Johnson, M. W. (2019). Cessation and reduction in alcohol consumption and misuse after psychedelic use. Journal of Psychopharmacology, 33(9), 1088–1101. https://doi.org/10.1177/0269881119845793

Goodwin, G. M., Aaronson, S. T., Alvarez, O., Arden, P. C., Baker, A., Bennett, J. C., Bird, C., Blom, R. E., Brennan, C., Brusch, D., Burke, L., Campbell-Coker, K., Carhart-Harris, R., Cattell, J., Daniel, A., DeBattista, C., Dunlop, B. W., Eisen, K., Feifel, D., . . . Malievskaia, E. (2022). Single-dose psilocybin for a treatment-resistant episode of major depression. New England Journal of Medicine, 387(18), 1637–1648. https://doi.org/10.1056/NEJMoa2206443

Griffiths, R. R., Johnson, M. W., Carducci, M. A., Umbricht, A., Richards, W. A., Richards, B. D., Cosimano, M. P., & Klinedinst, M. A. (2016). Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: A randomized double-blind trial. Journal of Psychopharmacology, 30(12), 1181–1197. https://doi.org/10.1177/0269881116675513

Johnson, M. W., Garcia-Romeu, A., Cosimano, M. P., & Griffiths, R. R. (2014). Pilot study of the 5-HT2AR agonist psilocybin in the treatment of tobacco addiction. Journal of Psychopharmacology, 28(11), 983–992. https://doi.org/10.1177/0269881114548296

Johnson, M. W., Garcia-Romeu, A., & Griffiths, R. R. (2017). Long-term follow-up of psilocybin-facilitated smoking cessation. American Journal of Drug and Alcohol Abuse, 43(1), 55–60. https://doi.org/10.3109/00952990.2016.1170135

Maisel, N. C., Blodgett, J. C., Wilbourne, P. L., Humphreys, K., & Finney, J. W. (2013). Meta-analysis of naltrexone and acamprosate for treating alcohol use disorders: When are these medications most helpful? Addiction, 108(2), 275–293. https://doi.org/10.1111/j.1360-0443.2012.04054.x

National Alliance on Mental Illness. (n.d.). Mental health by the numbers. https://www.nami.org/mhstats

Norring, S. A., & Spigarelli, M. G. (2024). The promise of therapeutic psilocybin: An evaluation of the 134 clinical trials, 54 potential indications, and 0 marketing approvals on ClinicalTrials.gov. Drug Design, Development and Therapy, 18, 1143–1151. https://doi.org/10.2147/DDDT.S443177

Ross, S., Bossis, A., Guss, J., Agin-Liebes, G., Malone, T., Cohen, B., Mennenga, S. E., Belser, A., Kalliontzi, K., Babb, J., Su, Z., Corby, P., & Schmidt, B. L. (2016). Rapid and sustained symptom reduction following psilocybin treatment for anxiety and depression in patients with life-threatening cancer: A randomized controlled trial. Journal of Psychopharmacology, 30(12), 1165–1180. https://doi.org/10.1177/0269881116675512

Wang, S.-M., Kim, S., Choi, W.-S., Lim, H. K., Woo, Y. S., Pae, C.-U., & Bahk, W.-M. (2024). Current understanding on psilocybin for major depressive disorder: A review focusing on clinical trials. Clinical Psychopharmacology and Neuroscience, 22(2), 222–231. https://doi.org/10.9758/cpn.23.1134

World Health Organization. (2023). Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression

Gene Editing and Synaptic Repair: Exploring CRISPR Applications in Restoring Neuronal Connectivity By Trishha Behal

Abstract

Numerous neurological diseases show synaptic dysfunction as a distinguishing pathology, such as Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis (ALS), spinal cord injury and autism spectrum disorder. As the synaptic plasticity and neuronal connectivity is progressively disrupted, cognitive function is impaired, motor function is impaired, and irreversible neuronal loss occurs; existing therapeutic options do not restore broken neural networks but only alleviate symptoms. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) gene-editing technologies have opened up novel opportunities for specifically targeting and correcting disease-associated genes causing synaptic degeneration.

This review discusses critically the molecular pathways involved in synaptic impairments and explores the therapeutic potential of CRISPR-Cas9, Cas12, Cas13, base editing and prime-Editing in reestablishing neuronal connectivity. The literature search was performed with peer-reviewed articles having been found exclusively in Google Scholar with a special focus on recent experimental and preclinical articles. The roles of key therapeutic targets, such as BDNF, MECP2, SHANK3, PSD95 (DLG4), APP, MAPT, APOE and SynGAP1 in synaptic plasticity, neuroregeneration and disease modification are analysed. Animal evidence reveals that CRISPR genome editing can increase dendritic spine density, promote the regeneration of axons, boost the transmission of synaptic signals, and restore behavioral function, albeit with great challenges in terms of off-target mutations, delivery to the brain, immune response and ethical concerns. The findings overall indicate that CRISPR-based gene editing could be a promising approach for precision medicine aimed at repairing neuronal networks and developing new therapeutic strategies for neurodegenerative and neurodevelopmental disorders in the future, but additional optimisation and clinical validation is needed before the approach can be widely used in therapeutic settings.

Keywords: CRISPR Gene Editing, Synaptic Repair, Neuronal Connectivity

Introduction

The prevalence of neurological diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), spinal cord injury (SCI), and autism spectrum disorder (ASD) accounts for over 3.4 billion people worldwide, the major cause of disability, and the second leading cause of death throughout the world (He et al., 2025). The characteristic pathological feature of all these conditions is synaptic dysfunction, which involves loss of neuronal communication, a decrease in synaptic plasticity and a progressive loss of neuronal connectivity. The relevance of synapse degeneration to learning, memory, cognition, motor control and in many cases irreversible neuronal death means that loss of synapses is often an early step in irreversible neuronal loss and is tightly coupled to loss of function. The clustered

regularly interspaced short palindromic repeats (CRISPR)-Cas genome-editing technology has revolutionized molecular neuroscience, allowing precise manipulation of disease-associated genes related to the formation of synapses, neurotransmission and neuronal survival. CRISPR systems, in comparison to traditional gene therapies, have important advantages like high editing specificity, programmability of the target sequence, and the ability to permanently fix pathogenic mutations, such as CRISPR-Cas9, base editing, and prime editing (Khan et al., 2026). Preliminary experiments have revealed editing efficiencies > 70-90% in specific models of neurons, leading to improvements in the expression of synaptic proteins, dendritic spine density and behavioral performance. This review critically discusses the molecular mechanisms of synaptic dysfunction and the existing evidence for the use of CRISPR-mediated gene editing as a therapeutic approach to re-establish neuronal connectivity. It also addresses key gene targets, experimental results, delivery systems, technical challenges and the future translational potential of precision neuroregenerative medicine.

Rationale

Recent advances in neuroscience have made profound strides, but these advances have led to a treatment strategy emphasizing symptom management, not repair of the underlying synaptic damage that leads to functional decline associated with neurological disorders. There is a strong correlation between synaptic loss and cognitive and motor dysfunction in a variety of diseases including Alzheimer's, Parkinson's and spinal cord injury, and thus the need for therapeutic interventions for the regeneration of synapses. A new approach is the use of CRISPR to edit genes critical for synaptic plasticity, neuronal survival and axonal regeneration, which has shown promise as a strategy (Wong et al., 2021). The review is significant because it critically appraises recent molecular and preclinical studies to assess the potential of these CRISPR technologies to restore neuronal connectivity and aid future clinical translation.

Objectives

1. To investigate the molecular bases of synaptic dysfunction and loss of neuronal connectivity in neurological diseases, focusing on those mechanisms that are genetic, such as those related to synaptic plasticity and repair.

2. To assess critically current CRISPR-based gene editing technology and potential approaches to restore synaptic function and neuronal regeneration, such as CRISPR-Cas9, base editing and prime editing.

3. To investigate experimental and preclinical studies of synaptic repair through CRISPR, considering the efficiency of CRISPR editing, therapeutic effects, technical problems and future perspectives in clinical neuroregenerative medicine.

Methodology

This systematic narrative literature review is a methodology that is used to critically analyze the use of CRISPR gene-editing technologies in synaptic repair and restoration of neuronal connectivity. The review relies only on peer-reviewed journal articles that have been cited in Google Scholar and therefore the evidence presented comes from high-quality, academic literature. Studies were identified using the following keywords: CRISPR-Cas9, gene editing, synaptic repair, neuronal connectivity, neuroregeneration, Alzheimer's disease, Parkinson's disease, spinal cord injury, base editing, prime editing. Newer publications (2019-2025) were prioritized for inclusion of recent advances in CRISPR technology and seminal studies published earlier were included if of scientific importance. Only articles in English, with full-text availability, and with experimental, preclinical or review-based evidence were included. The literature was critically reviewed and synthesized for the comparison of CRISPR platforms, target genes, editing efficiency, delivery systems and therapeutic outcomes across neurological disorders. Studies that were short on methodology detail and/or not scientifically robust were omitted to ensure the reliability of the review. Questions of ethics were related to the proper use of published research. No plagiarism, all information was paraphrased and sources correctly cited. The review also notes ethical issues related to the use of CRISPR technology, such as potential off-target genome editing, long-term safety, informed consent about the use of the technology for future clinical trials, and adherence to internationally agreed-on biomedical research and gene editing ethics.

Literature Review

Molecular Mechanisms of Synaptic Dysfunction and Neuronal Connectivity Loss

The role of synapses is to facilitate communication between two neurons in the form of a chemical or electric signal, and they are the structural basis for learning, memory, cognition and motor coordination. There are around 100 trillion synapses in the human brain in which the release of neurotransmitters from the presynaptic terminal activates postsynaptic receptors to modulate excitability of the neuron. Synaptic integrity is ensured by synchronous interactions between proteins of the synapse, between neurotransmitter receptors and intracellular signalling pathways. Disturbances of these mechanisms are known to impair neuronal connectivity and are an early pathological event in many neurodegenerative diseases (Südhof, 2021). The ability of synapses to enhance or impair their ability to carry out transmission as a function of their activity is a primary mechanism of synaptic plasticity, which is manifested as long-term potentiation (LTP) and long-term depression (LTD). LTP appears to be mostly mediated through N-methyl-D-aspartate (NMDA) receptors, which allow the entry of calcium ions and activate calcium/calmodulin-dependent protein kinase II (CaMKII) that strengthens synapses. On the other hand, too much LTD leads to the loss of dendritic spines and impairment of neuronal communication. Experimental studies indicate that about 40-60% of LTP can be lost by impairing NMDA receptor signaling, resulting in marked cognitive dysfunction in animal models

of Alzheimer's disease (AD) (Herring and Nicoll, 2016). There are multiple molecular pathways involved in maintaining and repairing synapses. Brain-Derived Neurotrophic Factor (BDNF) is responsible by binding to the TrkB receptor to improve neuronal survival, promote synaptic plasticity, and a decrease in the expression of BDNF has been correlated with a decrease in synaptic degeneration. Similarly, PSD-95 (DLG4) is a large postsynaptic scaffolding protein that stabilises NMDA receptors and structures excitatory synapses. SHANK3 helps regulate the organisation of the cytoskeleton in dendritic spines, while SynGAP1 is involved in inhibitory regulation of Ras signalling to maintain the balance of excitatory synapses. Mutations in SHANK3 and SynGAP1 have been reported to cause substantial decrease in dendritic spine density, typically 30-50% less than in healthy controls, and are associated with ASDs and intellectual disability (Monteiro and Feng, 2017). In fact, synaptic dysfunction is a common pathological feature of several neurological diseases. In Alzheimer's disease, soluble amyloid-β oligomers inhibit NMDA receptor signalling and hyperphosphorylated tau disrupts microtubules and leads to early loss of synapses prior to significant neuronal loss. Investigations following death estimate that up to 45% of the density of synapses is lost in individuals with AD, suggesting that this loss is more closely associated with cognitive decline than the amount of amyloid plaque. Post-mortem analyses show that there is up to 45% less density of synapses in individuals with AD compared to those without AD, with the amount of synapse loss being a better predictor of cognitive decline than the amount of amyloid plaque. Parkinson's disease involves loss of dopaminergic neurons in the substantia nigra that reduces the amount of dopamine available, resulting in decreased corticostriatal synaptic plasticity and motor function. Amyotrophic lateral sclerosis (ALS) is defined by glutamate-mediated excitotoxicity and dysfunction of the synaptic transmission system, while spinal cord injury (SCI) results in disruptions of the axons, inflammatory responses and failure of functional neuronal circuits. These results together show that the loss of neuronal connectivity is associated with the molecular disruption of synaptic proteins and signalling pathways, highlighting the importance of targeted therapeutic interventions such as gene editing by the CRISPR system to restore synaptic functions.

CRISPR Gene-Editing Technologies for Neurological Disorders

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) has transformed the way we are able to modify DNA and RNA sequences, allowing for precise and programmable genome editing to treat neurological diseases. CRISPR technologies are also adapted from the bacterial adaptive immune system and are now being explored as therapeutic agents to repair pathogenic mutations which cause the dysfunction of synapses, neuronal degeneration and impaired neural connectivity. CRISPR is more precise, scalable and can correct the genome permanently compared to conventional gene therapy (Doudna, 2020). The most popular platform is the CRISPR-Cas9, which uses a single-guide RNA (sgRNA) to guide the Cas9 endonuclease to a complementary DNA sequence on either side of a protospacer adjacent motif (PAM) sequence. Cas9 creates a site-specific doublestranded break in the DNA that is

repaired via non-homologous end joining (NHEJ) or homology-directed repair (HDR). The low proliferative capacity of neuronal cells limits the use of HDR, while NHEJ is more efficient and is the most preferred repair mechanism in post-mitotic neurons. Editing efficiencies have been reported in experimental studies ranging from 60 to >90% in neuronal cultures and animal models, depending on the gene being edited, the delivery vector and brain region (Anzalone et al., 2020). Newer CRISPR systems have broadened the therapeutic potential of CRISPR. Cas12a (Cpf1) is capable of recognising thymine-rich PAM sequences and introducing staggered DNA cleavages to enable large DNA insertions and multiplex genome editing while minimising off-target activity. Unlike Cas13 which targets RNA, however, and not DNA, Cas13 can only temporarily suppress the mutant transcripts without making permanent changes to the genome. This property is very desirable for dominant neurodegenerative disorders where reversible gene modulation is better desired (Abudayyeh et al., 2017). Recent advancements include base editing and prime editing that are able to address many of the limitations of traditional CRISPR-Cas9. Base editors consist of a catalytically inactive Cas protein and a deaminase enzyme that can be used to convert C•G to T•A or A•T to G•C without creating double-stranded DNA breaks. This strategy significantly decreases insertion-deletion (indel) mutations and shows editing efficiencies to be >80% in a number of experimental models. To enhance the genome-editing capabilities, prime editing also adds reverse transcriptase and a prime-editing guide RNA (pegRNA) to the genome-editing toolkit, allowing for targeted incorporation, deletion, and all 12 base changes without donor DNA. The efficiency of the Prime editing has been shown to be at 20-50 % correction rate, depending on the cell type and the locus of editing, with a low amount of unwanted genomic damage (Anzalone et al., 2019). Even with these developments in place, there still are a number of technical hurdles to be overcome. These effects, such as off-target editing, poor delivery into the brain and immune reactions against the Cas proteins, as well as low editing efficiency in differentiated neurons, still hinder the clinical translation of these methods. However, base and prime editing offer enhanced precision and safety compared with conventional CRISPR-Cas9, and it suggests a next generation of therapeutic platforms for neurological disorders to trigger targeted restoration of neuronal function.

CRISPR-Mediated Synaptic Repair and Neuroregeneration

Correcting genetic defects that result in defective neuronal connectivity via CRISPR gene editing has emerged as a potential therapeutic strategy to restore synaptic architecture and neuroregeneration. In contrast to traditional pharmacological approaches that target symptoms, CRISPR gene editing targets specific genes that play a key role in dendritic spine development, axonal growth, synaptic plasticity and neuronal survival - all of which are associated with disease. As a result, many preclinical models of neurological diseases have seen promising results with the use of CRISPR-based interventions for structural and functional recovery. Among the most important therapeutic benefits of the CRISPR editing is the restoration of postsynaptic dendritic spines, which are specialised structures that receive excitatory synaptic input. Among the most prominent studies is one that shows there is a strong association between

loss of dendritic spine density (DSD) and cognitive decline in Alzheimer's disease and in some neurodevelopmental disorders. SHANK3 and SynGAP1 mutations have been successfully corrected in a cell model using the CRISPR-Cas9 system, which restored dendritic spine morphology and raised synaptic density by 30-50%, which led to better neuronal communication and improved excitatory synaptic transmission (Mei et al., 2016). Likewise, higher BDNF levels in the dendrites result in dendritic branching and, via TrkB signalling, in maturation of synapses and long-term survival of the neurons in the brain, which is achieved by the use of CRISPR technology and the activation of the BDNF gene. CRISPR has also shown great promise when it comes to the regeneration of axons, especially after spinal cord injury (SCI). Inhibitory molecules (e.g., PTEN) inhibit re-growth of axons following CNS damage. Several studies have found that the regeneration of corticospinal tract axons after CRISPR-Cas9-induced knockout of PTEN is substantial, averaging 2-3-fold above the control. Increased expression of regeneration associated proteins, improved neuronal survival and partial reconstruction of damaged neural circuits (Liu et al., 2010) were associated with enhanced axonal regeneration. CRISPR also regulates genes for synaptogenesis and neurotransmitter signalling to stimulate the formation of new connections, in addition to repairing structures. The editing of genes coding for synaptic scaffolding proteins and neurotrophic signalling increases the efficiency of communication between neurons by making them more densely connected, both at the presynaptic and postsynaptic levels. These types of structural improvements correspond to functional recovery. In mouse models of AD and Rett syndrome, the genetic correction of disease-causing mutations using CRISPR led to considerable improvements in learning and memory, motor coordination and social behaviour. Electrophysiological measurements also showed that long-term potentiation (LTP) is restored, and that there is an increase in synaptic transmission and improvement of neuronal firing patterns, suggesting a recovery of functional synaptic plasticity (Abela et al., 2024). While the bulk of the evidence is anecdotal and from animal studies, the results are consistent, with all of these showing the potential to regenerate. The high editing efficiencies, a marked increase in dendritic spine density, as well as in axonal regeneration and behavioural outcomes, represent strong proof-of-concept that targeted genome editing is a promising approach to repair damaged neuronal networks and restore neurological function in clinical settings.

Target Genes Involved in Synaptic Repair

Successful application of CRISPR technology for neuroregeneration will primarily rely on choosing the right set of genes in play in the development and survival of synapses and how they contribute to neuronal plasticity. A number of genes have been found to be important for maintaining synaptic integrity, and these genes are now being explored to create genome-editing interventions of neurologic disorders (Akyuz et al., 2024). They have very different functions and can be used for different pathologies and therapeutic goals. One of the most widely researched neurotrophic proteins that control neuronal survival, dendritic growth and long-term potentiation (LTP) is called Brain-Derived Neurotrophic Factor (BDNF). Decreased BDNF

expression is consistently reported in Alzheimer's Disease, Parkinson's Disease and Huntington's Disease. Experimental CRISPR activation (CRISPRa) of BDNF has been shown to increase dendritic branching, expression of synaptic proteins and improvements in neuronal survival, suggesting that it has potential as a broad neuroprotective target (Kampmann, 2018). MECP2 is a gene that codes for methyl-CpG-binding protein 2, a protein that plays a critical role in regulating the epigenetic changes that are necessary for neuronal maturation and maintenance of synapses. Rett syndrome is a condition caused by mutations in the gene, MECP2, which leads to severe intellectual and motor disabilities. In preclinical models, the correction of pathogenic MECP2 mutations by the CRISPR approach has resulted in a restoration of neuronal morphology and an increase in synaptic transmission and notably, a marked improvement in neurological function. Therapeutic expression, however, needs to be carefully controlled since MECP2 deficiency and overexpression result in neurological abnormalities (Guan et al., 2022). The scaffolding proteins at synapses are crucial for the organisation of excitatory synapses, such as SHANK3 and PSD95 (DLG4). SHANK3 plays a role in maintaining the architecture of dendritic spines by anchoring neurotransmitter receptors to the actin cytoskeleton, while PSD95 anchors NMDA and AMPA receptors to postsynaptic membranes. SHANK3 loss-of-function mutations have a strong association with autism spectrum disorder and Phelan-McDermid syndrome. In experimental models, up to 30-50% of the expression of dendritic spines and the expression of synaptic proteins has been restored by CRISPR correction of SHANK3 mutations, while restoration of PSD95 expression increases the stability of synapses and improves excitatory neurotransmission. There are multiple therapeutic targets that are directly linked to the pathobiology of Alzheimer's disease. The Amyloid Precursor Protein (APP) gene is involved in the formation of amyloid-β peptides; the MAPT gene encodes the tau protein, which is involved in the stability of the microtubules. Mutant APP disruption via CRISPR results in a decrease in amyloid-β accumulation, and editing of MAPT results in a decrease in pathological tau phosphorylation and neurofibrillary tangle formation. Another major target is APOE, specifically the APOE4 allele, which is a strong risk factor for Alzheimer's disease, as it disrupts the clearance of amyloid, as well as promotes neuroinflammation (He et al., 2025). Recent studies of converting APOE4 to the protective APOE3 variant using CRISPR base-editing have shown reductions in amyloid pathology and increases in neuronal survival in laboratory models. Another important SynGAP1 target of synaptic plasticity is the regulation of Ras signalling and excitatory neurotransmission. When SynGAP1 is haploinsufficient, it leads to intellectual disability, epilepsy and autism spectrum disorders (Khan et al., 2026). Strong therapeutic potential was shown by experimental restoration of SynGAP1 expression, which improved the maturation of synapses and activity of neuronal networks. These genes, in general, are different in their mechanisms and disease specificity. While BDNF has wide-ranging neuroprotective activity, MECP2, SHANK3 and SynGAP1 have a specific activity for the monogenic neurodevelopmental disorders, and APP, MAPT and APOE are mainly active in the context of Alzheimer's disease pathology. Potential future CRISPR therapies will thus likely involve genetic

precision editing and personalised medicine strategies to maximize the ability to repair synapses and minimize unwanted effects on the genome.

Challenges, Ethical Considerations and Future Directions of CRISPR-Based Synaptic Repair

Although the ability of CRISPR to repair synapses and their role in treating neurological diseases has been shown to be of tremendous promise, there are a number of technical, biological and ethical obstacles that remain to be overcome before this technology can be used in clinical settings to treat these diseases. A major concern is off-target genome editing — the CRISPR system accidentally edits DNA sequences that are very similar to the desired sequence. These unwanted mutations can impact on critical genes, affect the performance of neurons, or raise the potential for the development of tumours. While high fidelity Cas9 variants and optimised guide RNA (gRNA) design has significantly minimized off-target activity, and advanced base and prime editing systems, editing specificity remains incomplete (Doudna, 2020). A challenge is the efficient delivery of CRISPR components across the blood brain barrier (BBB) which limits the delivery of larger biomolecules to the central nervous system. The most frequently used viral vectors at present are adeno-associated viruses (AAVs), which are highly efficient in transducing neurons and provide long-lasting gene expression (Wong et al., 2021). The DNA of AAV vectors, however, can only accommodate about 4.7 kb, making delivery of larger CRISPR systems difficult. Further, viral vectors can elicit immune reactions and carry a low risk of insertional mutagenesis. Nonviral delivery systems, such as lipid nanoparticles (LNPs), polymer-based nanoparticles, and extracellular vesicles, have higher safety profiles and reduced immunogenicity and are typically less efficient at editing and with less targeting precision in neuronal tissue. The host immune response is also a big problem. Cas proteins from bacteria like Staphylococcus aureus and Streptococcus pyogenes could be recognised by the human immune system, leading to antibody formation and activation of T cells and a diminished effect on treatment and risk of inflammation. Various strategies are being explored to reduce these immune-related complications, such as transient expression of CRISPR, engineering of Cas variants and immunomodulatory strategies. Ethically, there is a clear difference between somatic and germline genome editing. Somatic editing involves modification of non-reproductive cells, therefore would only impact the individual undergoing the process and not their offspring, which is the preferred option for neurological disorders. Unlike germline editing, which can lead to heritable genetic changes that can be passed down to future generations, there are serious ethical issues surrounding the safety, informed consent, equity, and potential misuse. (Marciano et al., 2022) As a result, clinical germline editing is strongly discouraged but carefully regulated somatic clinical applications for serious diseases are welcomed by international regulatory frameworks. Investigations into enhancing the precision of editing, next-generation delivery systems that penetrate the blood-brain barrier more effectively, and the safety of the editing over the long term - through large animal research and early phase clinical trials - continue to be a priority for future research. New developments in base editing, prime editing, AI-powered guide

RNA design and customised genomic treatments will further improve the precision of the treatment and minimise side effects (Abela et al., 2024). While the technology of synaptic repair using CRISPR is still largely in the preclinical phase, future studies indicate that further development and increased ethical scrutiny will be vital to the successful translation of this novel approach to safe and effective therapies for neurodegenerative and neurodevelopmental disorders.

Summary

This literature review highlights the potential of CRISPR-based gene editing technologies as a revolutionary solution for restoring neuronal connectivity by correcting genes associated with synaptic dysfunction and dysfunction. Preliminary findings show that CRISPR-Cas9, base editing and prime editing are able to increase dendritic spine density, promote axonal regeneration, increase synaptic plasticity and restore behavioural function in animal models of neurological disorders. Many genes, such as BDNF, MECP2, SHANK3, PSD95, APP, MAPT, APOE and SynGAP1, are good candidates for disease-specific therapeutic targets. Yet, potential problems like off-target mutations, blood-brain barrier delivery, immune reactions, and ethics need to be overcome to safely usher in a new era of routine clinical application of synaptic repair through CRISPR.

Conclusion

Overall, gene editing using CRISPR is a potential and still experimental approach for improving neuronal connectivity by specifically repairing molecular circuitries involved in synaptic dysfunction. The evidence presented above shows that disruption of synaptic plasticity, integrity of dendritic spines, neurotransmitter signalling and neuronal survival plays a significant role in functional impairment in Alzheimer's disease, Parkinson's disease, ALS, spinal cord injury and neurodevelopmental disorders. The DNA/RNA modification tools CRISPR-Cas9, Cas12, Cas13, base editing, and prime editing are more and more sophisticated tools to modify or target DNA or RNA including BDNF, MECP2, SHANK3, DLG4 (PSD95), APP, MAPT, APOE and SynGAP1. Preclinical studies have shown that in vivo modulation of these pathways can lead to changes in dendritic spine organisation, synaptic transmission, axon regeneration and behaviour with the potential to modulate disease mechanisms rather than just symptoms. The use of experimental models for clinical use is a challenge, however, due to off-target effects, restricted delivery to the blood brain barrier, immune responses, and editing efficiency as well as long-term safety concerns. Ethical issues, especially over heritable germline editing, also need to be closely monitored. In summary, while there are many exciting possibilities for the use of CRISPR-mediated synaptic repair in precision neuroregenerative medicine, careful validation, development of improved delivery methods and tools for the precise manipulation of editing and controlled clinical trials are necessary to make this a safe, effective therapy to restore neuronal connectivity.

Works Cited

Abela, L., Gianfrancesco, L., Tagliatti, E., Rossignoli, G., Barwick, K., Zourray, C., Reid, K.M., Budinger, D., Ng, J., Counsell, J. and Simpson, A., 2024. Neurodevelopmental and synaptic defects in DNAJC6 parkinsonism, amenable to gene therapy. Brain, 147(6), pp.2023-2037.

Abudayyeh, O.O. et al. (2017). RNA targeting with CRISPR-Cas13. Nature.

Akyuz, E., Aslan, F.S., Gokce, E., Ilmaz, O., Topcu, F. and Kakac, S., 2024. Extracellular vesicle and CRISPR gene therapy: Current applications in Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, and Huntington's disease. European Journal of Neuroscience, 60(8), pp.6057-6090.

Anzalone, A.V. et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature.

Anzalone, A.V. et al. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nature Biotechnology.

Anzalone, A.V. et al. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nature Biotechnology.

Arber, C. et al. (2021). Editing APOE4 to APOE3 using CRISPR-based approaches for Alzheimer's disease. Molecular Neurodegeneration.

Doudna, J.A. (2020). The promise and challenge of therapeutic genome editing. Nature.

Guan, L., Han, Y., Yang, C., Lu, S., Du, J., Li, H. and Lin, J., 2022. CRISPR-Cas9-mediated gene therapy in neurological disorders. Molecular Neurobiology, 59(2), pp.968-982.

He, C., Chen, B., Yan, C. and Zhou, X., 2025. Stem cell and CRISPR/Cas9 gene editing technology in Alzheimer’s disease therapy: from basic research to clinical innovation. Frontiers in genome editing, 7, p.1612868.

Herring, B.E. and Nicoll, R.A. (2016). *Long-Term Potentiation: From CaMKII to AMPA Receptor* Trafficking. Annual Review of Physiology.

Hsu, P.D., Lander, E.S. and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell.

Kampmann, M. (2018). CRISPR-based functional genomics for neurological disease research. Nature Reviews Neurology.

Khan, M.S., Zafar, I., Jamal, A., Bahwerth, F.S., Khan, S., Faizan, M., Khan, H., Noman, M., Okeibunor, K., Ahmad, A. and Zain Ul Abedin, M., 2026. CRISPR-Cas9 and next-generation gene editing strategies for therapeutic intervention of neurodegenerative pathways in Alzheimer’s disease: a state-of-the-art review. Acta Neurologica Belgica, pp.1-28.

Lim, C.K.W. et al. (2017). CRISPR-mediated gene correction restores neuronal function in neurological disease models. Nature Communications.

Liu, K. et al. (2010). PTEN deletion enhances the regenerative ability of adult corticospinal neurons. Nature Neuroscience.

Marciano, S., Ionescu, T.M., Saw, R.S., Cheong, R.Y., Kirik, D., Maurer, A., Pichler, B.J. and Herfert, K., 2022. Combining CRISPR-Cas9 and brain imaging to study the link from genes to molecules to networks. Proceedings of the National Academy of Sciences, 119(40), p.e2122552119.

Mei, Y. et al. (2016). Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. Nature.

Mei, Y. et al. (2016). Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. Nature.

Monteiro, P. and Feng, G. (2017). SHANK proteins: roles at the synapse and in autism spectrum disorder. Nature Reviews Neuroscience.

Moretti, P. and Zoghbi, H.Y. (2006). MeCP2 dysfunction in Rett syndrome and neurological disorders. Nature Reviews Neuroscience.

Naldini, L. (2015). Gene therapy returns to centre stage. Nature.

Südhof, T.C. (2021). The Cell Biology of Synapse Formation. Journal of Cell Biology.

Wang, D., Tai, P.W.L. and Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nature Reviews Drug Discovery.

Wong, P.K., Cheah, F.C., Syafruddin, S.E., Mohtar, M.A., Azmi, N., Ng, P.Y. and Chua, E.W., 2021. CRISPR gene-editing models geared toward therapy for hereditary and developmental neurological disorders. Frontiers in pediatrics, 9, p.592571.

Compact Object Mergers as Laboratories for Extreme Matter, Curved Spacetime, and Fundamental Physics By Dhruv Sharma

Abstract

With the detection of the first gravitational wave in 2015, and with the first detection of a binary neutron star merger with multi-messenger astronomy in 2017, compact object mergers have emerged as a leading probe of extreme astrophysics. This review discusses the key primary LIGO-Virgo-KAGRA Collaboration and related literature exemplifying one thesis, however limited: that compact binary mergers are laboratories for probes of strong-field gravity and dense matter as well as cosmology, with varying emphasis on each. The earliest observations constrained the equation of state of ultra-dense matter by measuring tidal deformability from neutron-star inspirals. Models consistent with the observations suggested relatively small neutron-star radii. The distance between these two values ranged between 11 and 13 km, depending on priors adopted, from the equation of state assumed, and whether observations from X-Rays and nuclear physics were mutually consistent. The second series of tests, using the compact-binary signals, tested predictions of general relativity in the strong-field, dynamical regime (and, with respect to Einstein's theory, the equality of light and gravity propagation speeds). Uncertainties in these null tests may arise from both the signal's amplitude and how the deviation is measured. For class three, the use of mergers as standard sirens within the distance ladder is limited by degeneracy between distance and inclination of the merger, low rate of electromagnetic counterpart detection, and lack of accurate galaxy catalogs. But combined with the first two classes, compact object merging gives very concrete observational prospects. At this point in time, GW170817 is the strongest evidence, as well as the clearest thorn in its side.

Background

Until recently, all astronomy made use of electromagnetic radiation, of all wavelengths from radio waves to X-rays to gamma-rays, as the messenger from the universe. Gravitational waves, however, are ripples of spacetime curvature first predicted by Einstein in 1916 because of his general theory of relativity, but they were not discovered until a century later. Because they interact so weakly with matter, gravitational waves propagate unaffected through events that do not emit electromagnetic radiation, including black hole mergers and neutron star mergers.

The main topics of this review are the remnants left when stars deplete their nuclear fuel and can no longer support themselves. A neutron star is a superdense object of a mass greater than that of the Sun, which is compressed into a sphere about the size of a city, with the density of atomic nuclei. A black hole is an object whose surface is so dense that not even light can escape it. General relativity predicts that if two such objects form a binary pair, they will spiral closer together as they lose energy to gravitational waves. The orbit speeds up as the objects come together, producing gravitational waves which have ever-increasing frequency and amplitude: a chirp. The frequency sweep is strongest at a combination of the two masses called the chirp mass; since this dominates the wave, it is almost always the best-measured property of

any detection. There is no unique frequency associated with the system, and the individual masses and spins are inferred from smaller features in the waveform. These signals are too weak to be detected by eye against the background noise in the data stream, and require matched filtering: cross-correlating data against banks of templates of theoretical waveforms calculated from general relativity and then identifying which template produces a statistically important correlation in the data stream (Krishnendu and Ohme). These sources may be classified as binary black holes, binary neutron stars, or binary neutron-star-black hole systems, with each source population illuminating a different corner of physics.

All of the sources cited here for the literature review mention some primary sources. A famous indirect detection was the discovery of the binary pulsar PSR B1913+16 by Hulse and Taylor in 1974, where a 30-year timing of orbital period decay agreed within 0.2% with the expected rate of energy loss due to gravitational radiation, winning them the 1993 Nobel Prize (Weisberg and Taylor). The September 2015 detection of gravitational waves from the merger of a 36 and 29 solar mass black hole at a distance of 1.3 billion light years (GW150914) with importance of greater than 5 sigma (Abbott et al., "Observation of Gravitational Waves"; Abbott et al., "Properties of the Binary Black Hole Merger GW150914") was followed by the August 2017 detection of the first binary neutron-star merger observed via gravitational waves and across the electromagnetic spectrum, inaugurating multi-messenger astronomy (Abbott et al., "GW170817: Observation of Gravitational Waves"). Since then the first such overlaps in the three types of thematic literatures reviewed below, indicating some overlap between the literatures.

This paper is not a standard review, but a survey of especially influential primary literature and associated reviews. These papers in science of neutron-star matter, strong field gravity, and standard-siren cosmology are cited repeatedly by other articles. The paper is therefore useful as a conceptual synthesis, but it does not try to provide a thorough coverage of other studies nor a formal statistical ranking of evidence. The question is not whether compact object mergers in each of these domains are valuable in their own right (this is well understood by the community), but whether, and how, the same events and the same data analysis pipelines really bring the two domains together. It is worth noting that this is an important caveat, because the review risks overweighting convergence if a few spectacular detections (most notably GW170817) are treated as a mature population.

My main thesis is that mergers of compact objects should be seen as a single physical laboratory rather than as a set of disjoint phenomena. I re-stress the idea because, from the perspective of using mergers as an observation platform, the scientific payoff varies markedly for different subjects within the field. They have enabled the strong-field gravity test, detailed formation and population of black-hole mergers, allowed dense-matter inference from the one neutron-star merger, and placed constraints on X-ray and nuclear physics. So far, they have not yet measured fast enough to yield a value of the Hubble parameter for standard siren cosmology. The rest of this paper reviews this discussion, including recent results in the equation of state of dense matter, general relativity in the strong field, and the cosmological expansion rate.

Compact Mergers as Probes of Ultra-Dense Matter

The purpose behind the papers in this first collection is to use the tidal interaction of inspiraling neutron stars to constrain the equation of state (EOS) of matter at supernuclear density. This regime is inaccessible to Earth-based experiments. This mechanism can be stated tidally, as it is what transforms the merger into a nuclear physics instrument. As these neutron stars approach each other the tidal force each invariant mass exerts on the other raises a tidal bulge on that star much as the Moon raises a tidal bulge on Earth. The degree to which a star can be deformed by this tidal force can be expressed by a dimensionless number, the tidal deformability, which is determined by the equation of state. The stiffness of the equation of state is, loosely speaking, a measure of the star's resistance to being compressed: an equation of state representing large stars which are weakly distorted has a high deformability. This deformation depletes some of the orbital energy faster than if the stars were point-like masses, so the gravitational wave phase is slightly advanced in the point-mass case. The offset can then be measured and used to determine the deformability and properties of matter that would otherwise remain out of reach.

The first results came from the LIGO-Virgo Collaboration's analysis of GW170817, where it was reported that the first binary neutron star merger had been detected and put an upper limit on the system's combined tidal deformability (Abbott et al., "GW170817: Observation of Gravitational Waves"). By taking a more advanced approach, an inference of about 222 under a uniform-mass prior was obtained, which changed the upper limit from 700 to 800 at this epoch (Abbott et al., "Properties of the Binary Neutron Star Merger GW170817"). When a joint analysis of the data with a common equation of state for both stars was performed, the common radius was found to be about 10.8 km with a range between 8.9–13.2 km (Abbott et al., "GW170817: Measurements of Neutron Star Radii"). These findings are important to this thesis, since they suggest a direct transition of waveform phase to nuclear physics.

It is important to note that the translation is not model-free. The tidal signal in GW170817 is inferred informative, but not precise enough to unambiguously determine the equation of state. The radius, however, changes when we vary the mass ratio, change the waveform model, change the spins, or change the parametrization of the tidal deformability. Going from 700 to 800 is not a crisis. Rather it is a warning that even headline numbers will change, as many choices in our analysis and source model are systematically revisited in this still young field. As illustrated by the literature on dense matter, the information that gravitational waves can provide is simultaneously powerful and limited. Although they are capable of placing a new constraint on the size of neutron stars, the information content requires a network of assumptions for interpretation and corroboration.

Several other works also self-consistently combine the gravitational-wave measurements with other astrophysical constraints. Most and colleagues argued that if the equation of state allows a two-solar-mass star (as required by the radio pulsar measurements) this severely constrains the possible range of radii for neutron stars. As such, these authors found that the radius of a 1.4-solar-mass neutron star lay between approximately 12.0 and 13.5 km (Most et

al.). This work is related to the thesis because it shows that the merger laboratory does not stand alone. However, its findings become stronger when combined with pulsar mass measurements, and are a limitation of the strongest version of the thesis. In the most constraining case, mass measurements from outside the merger, nuclear theory, and, eventually, X-ray observations are required. The merger provides an important pillar for multi-messenger inference, but is not a stand-alone laboratory, as the inferences are otherwise not as straightforward.

This limitation was highlighted by the second possible BNS merger discovery, GW190425, which had a mass of roughly 3.4 $M_\odot$, had no obvious EM counterpart, and no mechanism for breaking multi-messenger degeneracies with respect to counterparts. This made it less useful for much of the cross-domain analyses described in this review. It also raised questions about whether the Galactic sample is representative of the population observed by GW observatories. GW190425 thus gets in the way of a simple story about slow convergence. We can help, but the information that different detections yield is not the same. Some events sharpen the equation of state picture. Others simply test how complete the population picture is.

The third, most-complete, multi-messenger analysis includes gravitational-wave tidal deformability measurements, X-ray astrophysics from NASA's NICER, and nuclear physics. Both Miller et al. and Riley et al. independently measured the pulse profile of pulsar PSR J0030+0451 and claimed that concurrent mass-radius measurements were possible. Both experiments found a radius of about 13 km for a 1.4 solar mass star (Miller et al., "PSR J0030+0451"; Riley et al., "A NICER View of PSR J0030+0451"). The independent, pulse-profile based NICER analyses of the heavier PSR J0740+6620 in particular were important, because the star's mass is nearly two solar masses, requiring equations of state supported by higher central densities (Miller et al., "The Radius of PSR J0740+6620"; Riley et al., "A NICER View of the Massive Pulsar"). Raaijmakers et al. found canonical radii for neutron stars, combining and adapting GW170817, GW190425, NICER, radio timing, and nuclear theory, of about 12 km, depending on the equation of state parametrization. Similarly, using the reanalysis of NICER as well as modern insight on chiral effective-field theory, Rutherford et al. argued that the radius of a canonical neutron star is between 11.7 and 13.1 kilometers (Rutherford et al.).

Taken together the body of evidence is persuasive, but should not be overstated. The convergence at 11–13 km is one of the strongest results presented here, but is a convergence of overlapping credible intervals and compatible modeling frameworks, and is not a direct measurement in the sense that one would usually expect. It carries some genuine weight, but cannot exclude the possibility of a systematic error, or unify very soft equations of state, although it cannot ultimately distinguish between a diverse crust or core mixture and one containing some kind of exotic ingredient (hyperons, deconfined quark matter, etc.). The most honest conclusion is that compact mergers have made dense-matter physics observationally accessible, but have not fully solved the microphysical identification problem.

While most of the other work in the group looks at what happens inside the neutron star, the neutron-star merger produced a material that was ejected from the neutron stars at

GW170817 underwent r-process nucleosynthesis. A rapid sequence of neutron capture that has long been thought to produce about half of the elements heavier than iron, including the gold and platinum. The radioactive decay of freshly created material powered a kilonova, which was detected about eleven hours after the merger in the host galaxy NGC 4993. Pian et al. interpreted the color and decay rate of the kilonova, which was consistent with r-process material on the basis of the spectroscopic evidence of the kilonova, as direct evidence for the formation of heavy elements in a neutron-star merger. This essentially extends the merger laboratory, with the same event responsible for dense matter constraints also explaining the cosmic origin of heavy elements. However, the r-process interpretation is the most widely held. More detailed constraints on the element-by-element yields are more uncertain and depend on details of radiative-transfer modeling, ejecta geometry, and opacities.

The Compact Mergers as Tests of General Relativity

The second category of tests uses mergers to test general relativity in the strong-field, dynamical regime in a way that no previous test could. The classic tests of general relativity (bending of starlight, precession of Mercury, binary-pulsar timing tests described above) are tests of the weak-field limit or tests of systems moving slow with respect to the speed of light. On the other hand, as noted, during the final stages of a compact binary merger, the orbital separation is only a few kilometers, but relativistically large velocities and highly curved spacetime are involved. If GR fails anywhere in the region reachable from the Earth, it would be one of the first places where such failure would show.

The most relevant tests were on the third gravitational wave transient catalogue using the Collaboration's standardized battery of tests. Using multiple internal consistency checks (inspiral consistency test, merger-ringdown consistency test, anomalous dispersion test, additional polarizations), no statistically important deviation from general relativity was found (Abbott et al., "Tests of General Relativity with GWTC-3"). As a concrete example of what is made possible, an upper bound on the graviton mass can be obtained, less than about 1.27×10^{-23} eV/c², which is tighter than the corresponding Solar System limit. This result is relevant for the thesis of the paper, because all the equations which predict the mass, tidal deformability and distance of the binary also give rise to testable predictions.

It is important to stress, however, that these are null tests: the null result is only up to the sensitivity of the instruments, the quality of the signals observed and the specific alternative effect being tested. Therefore, one cannot claim that all modified gravity theories have been ruled out. Most of those tests in the catalog relate to the loudest and most easily parameterized tests of general relativity (as in binary black hole mergers), while the remainder, contrasting dense matter and bright siren cosmology, relate mostly to neutron-star mergers with electromagnetic counterparts. Thus, while the general relativity literature is closely related to these tests, it is less closely related to their evidence base.

The second branch of the field relates to the measurement technique. Krishnendu and Ohme review the field, and argue that gravitational waves provide qualitatively new ways to test

gravity, which can be divided into those due to the wave generation, those due to the wave propagation, and those arising from the merger remnant (Krishnendu and Ohme). The upshot of this section is that gravitational-wave tests are not merely quantitatively stronger versions of the Solar System or binary-pulsar tests; they test different regimes and different parts of the theory. The other implication is that, since all the tests above are cast in terms of parametrized deviations from general relativity, a theory producing deviations weak enough that they fall through the crack between parametrizations will be difficult or impossible to detect with any of the methods above.

GW170817 combined with the electromagnetic counterpart provides the most stringent limit of this kind. In this event, the gravitational-wave burst and the gamma-ray burst GRB 170817A were seen nearly simultaneously. The small difference in arrival time, 1.74 seconds after 130 million years of travel, constrains the difference between the speed of gravity and the speed of light to within a few parts in a thousand trillion (Abbott et al., "Gravitational Waves and Gamma-Rays"). GW170817 exemplifies the single merger event, with a variety of astrophysical applications: from constraining dense matter physics, testing gravity by measuring propagation speed and, by the distance and the redshift of the host galaxy, helping to constrain cosmology models.

Nevertheless, even though this is a strong propagation test, it must be said that this is not a complete test of general relativity. The travel time from GW170817 to Earth of the gravitational wave and photon emitted by the merger strongly constrains most modified gravity theories, particularly those predicting different propagation speeds for gravitational waves. However, it does not directly constrain other predictions of these theories, e.g. the detailed nonlinear black-hole ringdown dynamics or the properties of the interior of neutron-star matter. The conclusion is therefore balanced: compact mergers have made general relativity testable in a previously inaccessible regime, and all the studies we have critically reviewed are consistent with general relativity. There is confidence in this conclusion, but it is possible that more exotic signals might change this picture and test more deviations.

Compact Mergers as Cosmological Standard Sirens

All of the other works reviewed estimate the value for the Hubble constant, the rate of expansion of the universe, using mergers as standard sirens without utilizing the cosmic distance ladder. All the works reviewed conclude absolute distance can be determined from the gravitational-wave amplitude, a possibility provided by general relativity. Standard sirens can be measured directly from gravitational wave observation, which differ in their sources of uncertainty from both the early-universe measurement of the expansion rate from the cosmic microwave background and the late-universe measurement of the expansion rate from the distance to a distant supernovae determined by Cepheid-variable stars. Standard sirens can therefore provide an independent consistency check.

The theoretical basis of this method predates the detectors themselves. Schutz proposed in 1986 that gravitational-wave observations can determine the Hubble constant because the

gravitational-wave signature of a compact binary contains information about both the chirp mass and the luminosity distance (Schutz). The waveform amplitude measures the distance, just as the chirp mass determines the frequency sweep. This has the advantage that it can give the distance without the usual ladder of calibrated standard candles, but the distance is not enough. The redshift needs to be found. This may be either by direct detection of the host galaxy, or by association with a galaxy in the localization volume. It is at this stage that the method becomes observationally difficult.

The first observational measurement was from GW170817. The Collaboration and its partners viewed the gravitational wave signal, and located its optical counterpart in the host galaxy NGC 4993. They found $H_0 = 70$ km/s/Mpc (with an asymmetric uncertainty of roughly +12/−8) (Abbott et al., "A Gravitational-Wave Standard Siren Measurement"). The result was important because it does not depend on either the distance ladder or the cosmic microwave background. However, this uncertainty is too large to resolve the Hubble tension, and the main statistical problem is the distance-inclination degeneracy, whereby a relatively face-on system at one distance could be confused with a more inclined system at a different distance, and the gravitational-wave amplitude cannot always break the degeneracy.

Radio jet observations of GW170817 used to constrain the viewing angle showed the usefulness of electromagnetic data. The Hubble constant is now inferred to be 68.9 km/s/Mpc, with a much smaller uncertainty compared to the standard siren (Hotokezaka et al.). The ability to use electromagnetic data to constrain the Hubble constant makes mergers an even better cosmological tool. On the other hand, it shows a possible weakness of using mergers for cosmology. The best measurements are obtained if the gravitational-wave signal can be combined with a detailed electromagnetic model of a known nearby event. This means bright-siren measurements are informative, but not the most accurate ones, which require astrophysical modeling of the known events above.

Thus, the definition of a bright siren differs in the literature: a bright siren as GW170817 has an electromagnetic counterpart, from which both a unique host galaxy and a unique redshift can be inferred. Events like this are rare, and a dark siren does not have a counterpart, as is the case for most binary black hole mergers. The problem is statistical: the galaxies in the localization volume can give a possible redshift. A stack of many events strengthens the cosmological signal. In Fishbach et al. 's case, they excluded GW170817's host galaxy, and their result had a large uncertainty, but not less overall value being about 77 (the true error was not shown in their paper). This shows how well large catalogs and complete surveys will improve to make dark sirens competitive with standard sirens (cases that measure the distance and the electromagnetic redshift).

As such, this section provides the strongest test of the paper's unified-laboratory thesis. Strongly, standard sirens make compact mergers of cosmological tools. On the other hand, the literature on cosmology is less mature than that of dense matter or general relativity, statistically more difficult to study. For instance, the Planck Collaboration is able to determine the expanding rate of the Universe, based on the cosmic microwave background, as 67.4 kilometers per second

per megaparsec. Riess et al. get an expansion rate nearer to 73.0, yielding a five-sigma disagreement (Planck Collaboration; Riess et al.). Current standard siren distances do not have enough precision to discriminate between these poles. Their importance is not that they have solved the Hubble tension but that they are an independent line of inquiry that may well provide the resolution as more siren events are detected, and as the counterpart/galaxy catalog data grows.

Conclusion

Although the three literatures I have considered in this paper belong to the fields of nuclear physics, of gravitational theory, and of observational cosmology respectively, yet taken together they show an amazing structural isomorphism. A class of astrophysics observation, the gravitational wave (GW) events, shows directly that all three portions are possible. The gravitational waveform of a neutron star merger can be used to constrain the tidal deformability of the star, while the EM counterpart can be used to constrain the speed of gravity and the properties of the host galaxy. The distance-redshift relation can be used to make a standard siren measurement of the Hubble constant. This remains the strongest evidence for compact object mergers as a unified laboratory.

So the lesson is easily stated. While the unity is real, it is certainly incomplete. The detailed matter literature is moving forward. The last details will need NICER, gravitational wave observatories, the masses of radio pulsars, the modeling of kilonovae, and the nuclear theory behind these phenomena. The event sample is the largest in the relativity literature because all the available information is present in a single binary black hole merger. The only limitation for null tests is determined by the sensitivity of the measurements and the size of the deviation. By contrast, the standard-siren result is comparatively more conceptually straightforward (no distance ladder), but less precise, and does not resolve the Hubble tension. In addition, the same merger events may not necessarily carry all three domains along in the same way or to the same degree.

There are some common trends between different subfields, as the literatures are starting to interlace: gravitational-wave catalogs, electromagnetic counterparts, the era of modeling X-ray pulse profile, nuclear theory. Second, GW170817 is the most extreme case of and most cautionary tale for the field, but it also shows the huge potential of a multimessenger event to probe dense matter, gravity and cosmology. But it also shows that because it is currently the only such case, any general statements about compact mergers still rest on a very small number of especially informative observations. Third, each of these three disciplines has uncertainties: the composition of the densest stable stars, the relative validity of general relativity in the strong field, and the value of the Hubble constant.

Works Cited

Abbott, B. P., et al. "A Gravitational-Wave Standard Siren Measurement of the Hubble Constant." Nature, vol. 551, no. 7678, 2017, pp. 85–88.

Abbott, B. P., et al. "Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A." The Astrophysical Journal Letters, vol. 848, no. 2, 2017, L13.

Abbott, B. P., et al. "GW170817: Measurements of Neutron Star Radii and Equation of State." Physical Review Letters, vol. 121, no. 16, 2018, 161101.

Abbott, B. P., et al. "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral." Physical Review Letters, vol. 119, no. 16, 2017, 161101.

Abbott, B. P., et al. "GW190425: Observation of a Compact Binary Coalescence with Total Mass ~3.4 Solar Masses." The Astrophysical Journal Letters, vol. 892, no. 1, 2020, L3.

Abbott, B. P., et al. "Observation of Gravitational Waves from a Binary Black Hole Merger." Physical Review Letters, vol. 116, no. 6, 2016, 061102.

Abbott, B. P., et al. "Properties of the Binary Black Hole Merger GW150914." Physical Review Letters, vol. 116, no. 24, 2016, 241102.

Abbott, B. P., et al. "Properties of the Binary Neutron Star Merger GW170817." Physical Review X, vol. 9, no. 1, 2019, 011001.

Abbott, B., et al. "Tests of General Relativity with GWTC-3." Physical Review D, vol. 112, no. 8, 2025, 084080.

Fishbach, M., et al. "A Standard Siren Measurement of the Hubble Constant from GW170817 without the Electromagnetic Counterpart." The Astrophysical Journal Letters, vol. 871, no. 1, 2019, L13.

Hotokezaka, Kenta, et al. "A Hubble Constant Measurement from Superluminal Motion of the Jet in GW170817." Nature Astronomy, vol. 3, 2019, pp. 940–944.

Krishnendu, N. V., and Frank Ohme. "Testing General Relativity with Gravitational Waves: An Overview." Universe, vol. 7, no. 12, 2021, 497.

Miller, M. C., et al. "PSR J0030+0451 Mass and Radius from NICER Data and Implications for the Properties of Neutron Star Matter." The Astrophysical Journal Letters, vol. 887, no. 1, 2019, L24.

Miller, M. C., et al. "The Radius of PSR J0740+6620 from NICER and XMM-Newton Data." The Astrophysical Journal Letters, vol. 918, no. 2, 2021, L28.

Most, Elias R., et al. "New Constraints on Radii and Tidal Deformabilities of Neutron Stars from GW170817." Physical Review Letters, vol. 120, no. 26, 2018, 261103.

Pian, E., et al. "Spectroscopic Identification of r-Process Nucleosynthesis in a Possible Neutron Star Merger." Nature, vol. 551, no. 7678, 2017, pp. 67–70.

Planck Collaboration. "Planck 2018 Results. VI. Cosmological Parameters." Astronomy & Astrophysics, vol. 641, 2020, A6.

Raaijmakers, G., et al. "Constraints on the Dense Matter Equation of State and Neutron Star

Properties from NICER's Mass-Radius Estimate of PSR J0740+6620 and Multimessenger Observations." The Astrophysical Journal Letters, vol. 918, no. 2, 2021, L29.
Riess, Adam G., et al. "A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km/s/Mpc Uncertainty from the Hubble Space Telescope and the SH0ES Team." The Astrophysical Journal Letters, vol. 934, no. 1, 2022, L7.
Riley, Thomas E., et al. "A NICER View of PSR J0030+0451: Millisecond Pulsar Parameter Estimation." The Astrophysical Journal Letters, vol. 887, no. 1, 2019, L21.
Riley, Thomas E., et al. "A NICER View of the Massive Pulsar PSR J0740+6620 Informed by Radio Timing and XMM-Newton Spectroscopy." The Astrophysical Journal Letters, vol. 918, no. 2, 2021, L27.
Rutherford, N., et al. "Constraining the Dense Matter Equation of State with New NICER Mass-Radius Measurements and New Chiral Effective Field Theory Inputs." The Astrophysical Journal Letters, vol. 971, no. 1, 2024, L19.
Schutz, Bernard F. "Determining the Hubble Constant from Gravitational Wave Observations." Nature, vol. 323, no. 6086, 1986, pp. 310–311.
Weisberg, Joel M., and Joseph H. Taylor. "The Relativistic Binary Pulsar B1913+16: Thirty Years of Observations and Analysis." Binary Radio Pulsars, edited by F. A. Rasio and I. H. Stairs, ASP Conference Series 328, Astronomical Society of the Pacific, 2005, pp. 25–32.

Smart Manufacturing and Industrial Automation: A Comparative Review of AI Adoption in Kazakhstan and the United States By Zhanibek Alikhan

Abstract

The rapid development of artificial intelligence (AI), industrial automation, and Industry 4.0 technologies is transforming manufacturing worldwide by improving productivity, operational efficiency, and decision-making. Smart manufacturing integrates AI, the Internet of Things (IoT), robotics, cloud computing, and real-time data analytics to optimize production processes while reducing operational costs and equipment downtime. Although many countries have accelerated the adoption of these technologies, the pace of implementation differs depending on economic development, industrial infrastructure, and government policies. This paper compares the adoption of AI-driven smart manufacturing technologies in Kazakhstan and the United States, examining how each country utilizes industrial automation to improve manufacturing performance and competitiveness. Drawing on academic literature, industry reports, and government initiatives, the study evaluates the benefits, challenges, and future opportunities associated with digital transformation in manufacturing. The comparison demonstrates that while the United States leads in advanced AI integration and large-scale industrial automation, Kazakhstan has made significant progress through national digitalization programs and the gradual adoption of Industry 4.0 technologies. The findings suggest that continued investment in AI, workforce development, and smart manufacturing infrastructure will play an essential role in strengthening industrial competitiveness and supporting long-term economic growth.

Keywords: smart manufacturing, industrial automation, artificial intelligence, Industry 4.0, Kazakhstan, United States

1. Introduction

The manufacturing sector is undergoing one of the most significant technological transformations since the Industrial Revolution. Advances in artificial intelligence (AI), industrial automation, robotics, the Internet of Things (IoT), cloud computing, and data analytics have accelerated the transition toward smart manufacturing, often referred to as Industry 4.0. Unlike traditional manufacturing systems, smart factories continuously collect and analyze operational data, enabling machines to communicate, predict failures, optimize production schedules, and support real-time decision-making with minimal human intervention (Kusiak, 2018). According to the International Federation of Robotics (2024), more than 540,000 industrial robots were installed worldwide in 2023, highlighting the growing adoption of automation across manufacturing industries. At the same time, McKinsey & Company (2023) estimates that AI-powered manufacturing solutions can reduce equipment downtime by 30% to 50%, decrease maintenance costs by 10% to 40%, and improve overall productivity by 10% to 20%, making digital transformation a strategic priority for manufacturers worldwide.

The United States has emerged as one of the global leaders in smart manufacturing through initiatives such as Manufacturing USA, the National Institute of Standards and Technology (NIST) Smart Manufacturing Program, and substantial private investment in AI-driven production systems. Companies including General Electric, Siemens USA, Tesla, and Intel increasingly integrate predictive maintenance, industrial robotics, and digital twins into manufacturing operations to improve efficiency and competitiveness. Kazakhstan, while at an earlier stage of industrial digitalization, has also made measurable progress through the Digital Kazakhstan initiative and national Industry 4.0 programs aimed at modernizing key industries such as mining, metallurgy, oil and gas, and manufacturing. Studies indicate that digital technologies, including AI, IoT, and industrial automation, are gradually being introduced into Kazakhstan's industrial sector, particularly among large enterprises seeking to improve operational efficiency and global competitiveness (Dikhanbayeva et al., 2022). However, compared with the United States, adoption remains uneven due to differences in technological infrastructure, investment capacity, workforce readiness, and industrial maturity.

This paper compares the adoption of AI-driven smart manufacturing technologies in Kazakhstan and the United States by reviewing contemporary academic literature, government reports, and industrial case studies. It examines how artificial intelligence, automation, and Industry 4.0 technologies improve manufacturing performance while also identifying the technical and economic challenges associated with digital transformation. The paper argues that although the United States currently demonstrates a higher level of technological integration, Kazakhstan has established a promising foundation for future industrial modernization. Continued investment in digital infrastructure, engineering education, and intelligent manufacturing systems will be essential for narrowing the technological gap and strengthening industrial competitiveness in both domestic and international markets.

2. Literature Review

The concept of smart manufacturing refers to the integration of advanced digital technologies into production systems to improve efficiency, flexibility, and decision-making. Unlike conventional manufacturing, which relies heavily on manual monitoring and scheduled maintenance, smart manufacturing combines artificial intelligence (AI), the Internet of Things (IoT), cloud computing, industrial robotics, digital twins, and big data analytics to create intelligent production environments capable of self-monitoring and continuous optimization (Kusiak, 2018). These technologies enable manufacturers to collect and analyze large volumes of operational data in real time, allowing production systems to respond automatically to changing conditions. According to the World Economic Forum (2023), manufacturers implementing Industry 4.0 technologies have reported productivity improvements of 15% to 30%, while digital factories have reduced production costs by up to 20% through better resource utilization and process optimization. Similarly, Frank et al. (2019) argue that smart manufacturing should be viewed as a comprehensive transformation of industrial operations rather than simply the

adoption of new technologies, requiring changes in organizational strategy, workforce skills, and business processes.

Artificial intelligence has become one of the primary drivers of smart manufacturing because it enables predictive analytics, intelligent automation, and data-driven decision-making across production systems. Machine learning algorithms can analyze sensor data from industrial equipment to predict failures before they occur, reducing unplanned downtime and improving equipment reliability. Bousdekis et al. (2021) found that AI-based predictive maintenance significantly improves asset utilization while reducing maintenance costs and increasing production efficiency. Likewise, Lee et al. (2018) emphasize that AI enhances manufacturing performance by optimizing production scheduling, quality control, energy consumption, and supply chain management simultaneously. According to Deloitte (2024), more than 70% of manufacturing executives consider AI to be essential for maintaining long-term competitiveness, while manufacturers investing in AI report measurable improvements in operational efficiency, product quality, and decision-making speed. These findings demonstrate that AI is no longer viewed as an experimental technology but as a strategic capability supporting modern industrial competitiveness.

Despite widespread recognition of its benefits, the adoption of smart manufacturing technologies varies considerably across countries. The United States has established one of the world's most advanced manufacturing innovation ecosystems through initiatives such as Manufacturing USA, the National Institute of Standards and Technology (NIST), and extensive private-sector investment in industrial AI, robotics, and digital manufacturing. In contrast, Kazakhstan remains in an earlier phase of Industry 4.0 implementation, although recent studies indicate steady progress through the Digital Kazakhstan program and increasing adoption of AI, IoT, and industrial automation within mining, metallurgy, and manufacturing enterprises (Dikhanbayeva et al., 2022). Research on Kazakhstan's manufacturing sector suggests that the primary barriers to wider implementation include limited digital infrastructure, shortages of highly qualified engineering specialists, and relatively high investment costs for advanced manufacturing technologies. Nevertheless, scholars argue that continued modernization efforts and international technology partnerships provide significant opportunities for accelerating industrial digitalization and improving the global competitiveness of Kazakhstan's manufacturing sector (Babkin et al., 2024).

3. Comparative Analysis of AI Adoption in Manufacturing: Kazakhstan and the US

Although both Kazakhstan and the United States recognize artificial intelligence as a key driver of industrial development, the scale and maturity of AI adoption differ considerably. The United States is among the global leaders in smart manufacturing, supported by strong collaboration between government, industry, and research institutions. Programs such as Manufacturing USA and the National Institute of Standards and Technology (NIST) promote the development and implementation of advanced manufacturing technologies, including AI, industrial robotics, digital twins, and cyber-physical systems. According to the International

Federation of Robotics (2024), the United States ranked among the world's largest markets for industrial robots, with manufacturing sectors such as automotive, electronics, and aerospace accounting for the majority of new installations. Furthermore, McKinsey & Company (2023) reported that nearly 55% of organizations globally have adopted AI in at least one business function, with manufacturing representing one of the fastest-growing application areas. American manufacturers increasingly employ AI for predictive maintenance, automated quality inspection, supply chain optimization, and production planning, allowing them to improve productivity while reducing operational costs.

Kazakhstan has made significant progress in industrial digitalization over the past decade, although AI adoption remains concentrated in large enterprises rather than across the entire manufacturing sector. The government's Digital Kazakhstan initiative, launched to accelerate technological modernization and economic diversification, has encouraged investments in industrial automation, digital infrastructure, and Industry 4.0 technologies. According to Dikhanbayeva et al. (2022), sectors such as mining, metallurgy, oil and gas, and textile manufacturing have begun integrating IoT devices, automation systems, and data analytics to improve production efficiency and resource management. However, compared with the United States, many small and medium-sized manufacturing enterprises continue to face challenges related to limited financial resources, insufficient digital infrastructure, and shortages of engineers with expertise in AI, robotics, and industrial data analytics. Researchers also note that successful implementation of smart manufacturing requires not only technological investment but also workforce training, organizational change, and stronger collaboration between universities, industry, and government agencies (Babkin et al., 2024).

In the face of these differences, both countries share similar long-term objectives of improving industrial productivity, competitiveness, and sustainability through digital transformation. The United States currently demonstrates greater technological maturity due to decades of investment in research, innovation, and advanced manufacturing ecosystems. Kazakhstan, however, possesses substantial opportunities to accelerate AI adoption by leveraging international partnerships, expanding engineering education, and continuing national digitalization programs. Rather than replicating the American model directly, Kazakhstan can adopt technologies that address the specific needs of its industrial sectors while gradually expanding the use of AI-driven manufacturing solutions. This comparative analysis suggests that although the two countries differ in their current level of technological development, both recognize smart manufacturing as a strategic priority for strengthening industrial competitiveness in the global economy.

4. Benefits and Challenges of Smart Manufacturing

The adoption of AI and industrial automation offers substantial benefits for manufacturing organizations by improving operational efficiency, product quality, and long-term competitiveness. One of the most significant advantages is the ability to make data-driven decisions in real time. AI-powered systems continuously analyze information collected from

sensors, machines, and production lines, enabling manufacturers to identify inefficiencies, optimize workflows, and respond quickly to unexpected disruptions. Predictive maintenance has become one of the most valuable applications of AI because it allows companies to detect equipment failures before they occur, reducing downtime and maintenance costs. According to McKinsey & Company (2023), predictive maintenance can reduce equipment downtime by 30% to 50% while lowering maintenance costs by 10% to 40%. In addition, AI-supported quality inspection systems improve product consistency by identifying manufacturing defects with greater speed and accuracy than traditional manual inspections (Lee et al., 2018). These improvements contribute directly to higher productivity, lower operating costs, and greater customer satisfaction.

Despite these advantages, implementing smart manufacturing technologies presents several technical and organizational challenges. One of the primary obstacles is the high initial investment required for AI software, industrial robots, IoT sensors, cloud infrastructure, and employee training. While large corporations often possess the financial resources to modernize production facilities, small and medium-sized manufacturers may struggle to justify these investments despite their long-term benefits. Another challenge involves workforce readiness. As manufacturing becomes increasingly digital, engineers and technicians require new competencies in artificial intelligence, data analytics, cybersecurity, and industrial automation. According to the World Economic Forum (2023), the shortage of digital and engineering skills remains one of the most significant barriers to successful Industry 4.0 implementation worldwide. Cybersecurity has also become a growing concern because connected manufacturing systems generate and exchange large volumes of sensitive operational data, making them attractive targets for cyberattacks.

These challenges are particularly relevant for Kazakhstan as the country continues expanding its industrial digitalization efforts. Although national initiatives such as Digital Kazakhstan have accelerated the adoption of Industry 4.0 technologies, further progress depends on continued investment in digital infrastructure, engineering education, research partnerships, and innovation ecosystems. At the same time, the United States continues to face its own challenges, including cybersecurity risks, workforce shortages, and the complexity of integrating AI into existing manufacturing systems. Therefore, successful implementation of smart manufacturing depends not only on technological innovation but also on effective government policies, collaboration between universities and industry, and the development of a highly skilled engineering workforce capable of supporting long-term industrial transformation.

5. Discussion

The comparison between Kazakhstan and the United States demonstrates that artificial intelligence and industrial automation have become essential components of modern manufacturing rather than optional technological innovations. The reviewed literature consistently shows that AI improves production efficiency through predictive maintenance, intelligent quality control, real-time monitoring, and optimized production planning. While both

countries recognize the strategic importance of smart manufacturing, their approaches differ because of variations in technological infrastructure, research capacity, investment levels, and industrial maturity. The United States benefits from a well-established innovation ecosystem supported by government initiatives, leading universities, and private-sector investment, allowing manufacturers to integrate advanced AI technologies on a large scale. Kazakhstan, in contrast, is building its smart manufacturing capabilities through national digitalization strategies and gradual modernization of key industries, particularly mining, metallurgy, and manufacturing. Although the pace of implementation is slower, current initiatives indicate a strong commitment to long-term industrial transformation.

The literature also suggests that technology alone is insufficient to achieve successful digital transformation. Sustainable adoption of smart manufacturing depends on workforce development, engineering education, cybersecurity, and collaboration between governments, universities, and industry. For Kazakhstan, strengthening partnerships with international technology companies and expanding AI-related engineering education could accelerate the adoption of Industry 4.0 technologies while improving industrial competitiveness. At the same time, the United States continues addressing challenges related to cybersecurity, workforce shortages, and the responsible integration of increasingly complex AI systems. Overall, the evidence indicates that countries investing not only in advanced technologies but also in human capital and innovation ecosystems will be better positioned to compete in the future of global manufacturing.

6. Conclusion

This paper compared the adoption of artificial intelligence and industrial automation in the manufacturing sectors of Kazakhstan and the United States, highlighting both the opportunities and challenges associated with smart manufacturing. The literature demonstrates that AI-driven technologies such as predictive maintenance, industrial robotics, IoT, and data analytics significantly improve productivity, operational efficiency, product quality, and decision-making. The United States currently leads in the large-scale implementation of these technologies through strong investment, research institutions, and advanced manufacturing initiatives. Kazakhstan, while at an earlier stage of digital transformation, has made meaningful progress through the Digital Kazakhstan program and the gradual adoption of Industry 4.0 technologies across major industrial sectors.

The comparison shows that future industrial competitiveness will depend not only on access to advanced technologies but also on the ability to develop a skilled engineering workforce, modern digital infrastructure, and effective collaboration between government, academia, and industry. Continued investment in AI, automation, and engineering innovation will enable both countries to strengthen manufacturing resilience and respond more effectively to the demands of an increasingly digital global economy. For Kazakhstan in particular, expanding the implementation of smart manufacturing technologies represents an important opportunity to modernize industry, diversify the economy, and increase international competitiveness.

Works Cited

Babkin, A. V., Tashenova, L. V., Mamrayeva, D., & Sarsenbayev, M. (2024). International experience of adaptation and application of digital ecosystems by industrial enterprises and clusters. BIO Web of Conferences, 64, 04048. https://doi.org/10.1051/bioconf/20246404048

Bousdekis, A., Magoutas, B., Apostolou, D., & Mentzas, G. (2021). Review, analysis and synthesis of prognostic-based decision support methods for predictive maintenance in manufacturing. Journal of Intelligent Manufacturing, 32(3), 687–711. https://doi.org/10.1007/s10845-020-01672-1

Dikhanbayeva, D., Aitzhanova, M., Shehab, E., Turkyilmaz, A., & Nurdauletov, K. (2022). Analysis of textile manufacturing SMEs in Kazakhstan for Industry 4.0. Procedia CIRP, 107, 1438–1443. https://doi.org/10.1016/j.procir.2022.05.162

Deloitte. (2024). State of generative AI in the enterprise. Deloitte Insights. https://www2.deloitte.com

Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. International Journal of Production Economics, 210, 15–26. https://doi.org/10.1016/j.ijpe.2019.01.004

International Federation of Robotics. (2024). World Robotics 2024: Industrial robots. https://ifr.org

Kusiak, A. (2018). Smart manufacturing. International Journal of Production Research, 56(1–2), 508–517. https://doi.org/10.1080/00207543.2017.1351644

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manufacturing Letters, 18, 20–23. https://doi.org/10.1016/j.mfglet.2018.09.002

McKinsey & Company. (2023). The state of AI in 2023: Generative AI’s breakout year. https://www.mckinsey.com

National Institute of Standards and Technology. (2020). Manufacturing USA and smart manufacturing initiatives. https://www.nist.gov

Organisation for Economic Co-operation and Development. (2021). The digital transformation of SMEs. OECD Publishing. https://www.oecd.org

World Economic Forum. (2023). Global Lighthouse Network: Unlocking sustainability through digital manufacturing. https://www.weforum.org

World Bank. (2023). Digital development in Kazakhstan: Opportunities for industrial transformation. World Bank. https://www.worldbank.org

Effect of Microbial Necromass Deactivation Method on Radish Growth in Lunar Regolith Simulant By Shonik Shah

Abstract

Sustained human presence on the Moon will require bioregenerative agricultural systems capable of producing food using in-situ resources. Lunar regolith is inherently sterile and lacks organic carbon, microbial nutrient cycling, and bioavailable nitrogen and phosphorus, severely constraining plant growth. One strategy for overcoming these limitations is the incorporation of microbial necromass, the nonliving residual biomass of microbial cells, which contains structurally complex organic compounds capable of acting as slow-release nutrient reservoirs while improving soil structure and water retention. In lunar habitats, microbial biomass would naturally accumulate in waste recycling systems, composting units, and other bioregenerative life-support bioreactors that process human waste and plant residues. When microbial communities die or are intentionally sterilized within these life-support systems, they produce necromass that could potentially be recycled as a soil amendment for extraterrestrial agriculture.

This study investigated whether sterile microbial necromass can enhance plant growth in lunar regolith simulant and whether the microbial inactivation pathway alters its effectiveness as a soil amendment. Radish (*Raphanus sativus*) was cultivated in three regolith treatments: untreated regolith (control), regolith amended with heatkilled microbial necromass, and regolith amended with ethanol-killed microbial necromass, with necromass incorporated at 15% by dry mass.

Across replicate trials, necromass amendments substantially improved plant performance relative to untreated regolith over the 20-day growth period. On average across trials, mean plant height increased from **6.3 cm in the control treatment to 14.2 cm with heat-killed necromass and 14.1 cm with ethanol-killed necromass**, representing approximately a **120% increase in growth**. Leaf production increased from an average of **5 leaves in the control to 36 leaves in the heat-treated system and 22 leaves in the ethanol treatment**, while necromass amended plants maintained higher physiological health scores and reduced wilting. Microbial necromass significantly increased plant productivity, with mean fresh biomass rising from 0.21 g in the control to 0.68 g (ethanoltreated) and 1.03 g (heat-treated), confirming enhanced nutrient availability in regolith systems. Overall, heat-killed necromass produced the strongest growth response, generating slightly greater plant height, fresh biomass and substantially higher leaf production than ethanol-killed necromass, although both treatments significantly improved plant performance relative to untreated regolith. These findings demonstrate that recycled microbial necromass, particularly heat-inactivated biomass, can function as a biologically derived soil amendment for lunar agriculture, supporting nutrient recycling and sustainable crop production within closed-loop life-support systems. This approach directly addresses the fundamental limitation of lunar regolith—its severe nutrient deficiency—by enabling the in situ conversion of waste

microbial biomass into a functional growth resource, providing a novel closed-loop solution that transforms waste streams into usable biological inputs and eliminating dependence on Earth-supplied materials. Importantly, microbial biomass is continuously generated as a byproduct of bioregenerative life-support systems, making it a self-renewing resource stream that can potentially sustain long-term agricultural cycles in closed extraterrestrial habitats in a novel manner.

Introduction

Sustained human presence on the Moon requires agricultural systems capable of producing food using in-situ resources while minimizing dependence on resupply from Earth. Lunar regolith, however, differs fundamentally from terrestrial soil. Formed through billions of years of meteoritic bombardment and physical weathering rather than biological processes, regolith lacks organic matter, microbial ecosystems, and biologically mediated nutrient cycling. Although it contains mineral elements such as silicon, aluminum, calcium, and iron, it is severely deficient in organic carbon and bioavailable nitrogen and phosphorus, making it inherently unsuitable for sustained plant growth.

Recent experimental work has demonstrated that plants can germinate in authentic lunar regolith samples returned during the Apollo missions. However, plant development remains strongly constrained in these substrates. In a landmark study, *Arabidopsis thaliana* was successfully germinated in lunar regolith but exhibited reduced biomass accumulation, physiological stress, and altered gene expression associated with nutrient limitation and abiotic stress (Ferl & Paul, 2022). These findings indicate that although regolith can physically support germination, the absence of organic nutrient pools and microbial nutrient cycling prevents long-term plant productivity. Several strategies have been proposed to overcome these limitations, including the addition of terrestrial soil amendments, synthetic fertilizers, composted plant residues, or engineered microbial communities. However, many of these approaches rely on materials that would require continual importation from Earth or large-scale biological infrastructure that may be difficult to maintain in early extraterrestrial habitats. Consequently, current research in space agriculture increasingly emphasizes closed-loop resource recycling systems capable of converting biological waste streams into usable nutrients within life-support architectures.

One potential nutrient source within closed-loop life-support systems is **microbial necromass**, the residual non-living biomass produced when microbial cells die or are intentionally inactivated. In terrestrial soils, the continuous turnover of microbial communities generates necromass composed of cellular structural components including proteins, lipids, polysaccharides, and amino-sugar–containing cell wall fragments. Soil biogeochemistry studies show that these residues form a major fraction of stable soil organic matter and represent an important reservoir of plant-available carbon and nitrogen (Liang et al., 2019). Necromass accumulation is commonly identified through amino sugar biomarkers such as muramic acid and glucosamine, which originate from bacterial and fungal cell walls and

persist in soil long after microbial death. These compounds contribute to the formation of relatively stable organic matter pools that gradually release nutrients as they decompose. In addition to acting as nutrient reservoirs, necromass-derived molecules can interact with mineral surfaces through **organo-mineral associations**, a process that stabilizes organic carbon while increasing nutrient retention and water-holding capacity within soils. Because lunar regolith is composed largely of fine mineral particles but lacks organic matter, the introduction of necromass could provide both a biochemical nutrient source and an organic matrix capable of improving substrate structure. These characteristics suggest that microbial necromass may function as a biologically generated soil amendment capable of enhancing the fertility of regolith-based growth substrates used in extraterrestrial agriculture.

Future lunar habitats are expected to operate as **Closed Ecological Life Support System (CELSS)**, incorporating microbial bioreactors, composting systems, and waste-processing units that recycle plant residues, food waste, and human waste into usable resources. Within these systems, microbial populations will continually grow and die as part of normal metabolic turnover, leading to the accumulation of substantial quantities of microbial necromass as an inevitable byproduct of life-support operations. Repurposing this necromass as a regolith amendment could therefore provide a nutrient source that is internally generated within the habitat rather than transported from Earth, making it a potentially sustainable strategy for extraterrestrial agriculture. Recycling microbial necromass in this manner aligns with the principle of **in-situ resource utilization (ISRU)**, a key strategy in planetary exploration that seeks to convert locally available materials into usable resources for sustaining human activity beyond Earth. Establishing such recycling pathways represents the early development of **artificial biogeochemical cycles**, which are essential for maintaining stable nutrient flows in closed ecological systems.

Despite extensive research on microbial necromass dynamics in terrestrial soils, its potential application as an amendment for extraterrestrial regolith remains largely unexplored. Furthermore, an important factor influencing necromass functionality may be the method by which microbial cells are inactivated prior to incorporation into growth substrates. In extraterrestrial habitats, microbial biomass will frequently undergo intentional sterilization to comply with planetary protection protocols designed to prevent forward contamination of celestial bodies. Consequently, microbial necromass produced within life-support systems will likely originate from sterilized microbial biomass rather than natural microbial death. The biochemical composition and nutrient availability of necromass depends on the method used to deactivate microbial cells. Thermal sterilization can denature proteins, disrupt cell walls, and accelerate molecular breakdown, potentially increasing the accessibility of intracellular nutrients. In contrast, ethanol sterilization primarily permeabilizes cellular membranes while preserving much of the internal cellular structure, which may influence the rate at which nutrients are released into the surrounding substrate.

Therefore, this study investigates not only whether microbial necromass can enhance plant growth in lunar regolith simulant but also evaluates how different microbial inactivation

pathways influence its effectiveness as a soil amendment. Two inactivation methods—thermal sterilization and ethanol sterilization— were selected because they represent sterilization approaches that are both technologically feasible within lunar habitats. Thermal sterilization could occur through high-temperature waste processing units or thermal treatment systems used to sanitize biological waste streams, while ethanol-based sterilization could be implemented through chemical disinfection protocols routinely used to control microbial contamination in closed spacecraft environments. Both approaches align with sterilization procedures already used in spacecraft operations and could be integrated into life-support waste recycling systems within lunar habitats .

Alternative microbial inactivation methods exist, including ultraviolet (UV) irradiation, autoclave-based steam sterilization, and chemical oxidizing agents such as chlorine or hydrogen peroxide. However, these methods present limitations for necromass recycling in extraterrestrial agriculture. UV sterilization is primarily effective for surface decontamination and has limited penetration through dense biomass, reducing its effectiveness for bulk microbial substrates (Kowalski, 2009). Steam autoclaving requires substantial water resources and high-pressure systems, which may be impractical in early lunar habitats where water availability and infrastructure are constrained. Strong oxidizing chemical disinfectants such as chlorine or hydrogen peroxide can degrade organic molecules and generate reactive by-products that may reduce the nutritional value of microbial biomass intended for soil amendment (Rutala & Weber, 2016). In contrast, thermal inactivation and ethanol sterilization can effectively deactivate microbial cells while largely preserving the organic molecular components that constitute necromass, making them suitable.

Experimental Design

The experiment was conducted using a controlled indoor plant growth setup in a home environment to allow continuous daily monitoring under stable conditions. Plants were grown under natural daylight with an approximately **16 h light / 8 h dark photoperiod** at an ambient temperature of ~**28 °C**. This temperature range was selected because radish germination and early vegetative metabolism occur optimally between **20–30 °C**, where enzymatic activity, respiration rates, and cellular division processes supporting seedling establishment are maximized. Extended photoperiod conditions approximate controlled environment agriculture systems used in space crop production, where increased light availability enhances photosynthetic carbon assimilation and accelerates early biomass accumulation.

Three substrate treatments were established:

(1) **control-100 % regolith simulant**,

(2) **regolith amended with heat-killed microbial necromass**,

(3) **regolith amended with ethanol-killed microbial necromass**.

Microbial necromass will be derived from a standardized compost tea batch and rendered non-viable via heat or ethanol treatment to test whether the mode of necromass inactivation influences plant growth responses. Heat-killed necromass will be produced via boiling, followed by filtration, rinsing with sterile distilled water, and air-drying. Ethanol-killed necromass will be prepared using a 70% ethanol treatment, with repeated rinsing to eliminate residual solvent prior to drying. Necromass was prepared using a standardized protocol from a single homogenized batch of compost tea to ensure consistency.

To confirm effective microbial inactivation and prevent residual viable microorganisms from affecting plant growth, sterility of both necromass treatments will be assessed using a closed-container moisture-activated regrowth assay. Subsamples will be placed in clean transparent containers, moistened with sterile distilled water, and loosely sealed to allow limited gas exchange. Containers will be incubated at room temperature (22–28 °C) for five days and examined daily for indicators of microbial activity, including turbidity, gas production, biofilm formation, visible fungal growth, and odor changes. Absence of these indicators will indicate successful inactivation. A negative control with distilled water and a positive control with untreated compost residue will verify absence of contamination and suitability of growth conditions. Differences in regrowth, if observed, will contextualize potential effects on necromass biochemical integrity rather than microbial viability.

In amended treatments, microbial necromass constituted **15 % of the substrate by dry mass**, producing a mixture of **204 g regolith simulant and 36 g necromass per pot (240 g total)**. This proportion was selected to introduce a biologically meaningful organic carbon and nitrogen input while maintaining the **mineralogical dominance of the regolith matrix (85 %)**. Preserving regolith as the primary structural component more realistically represents extraterrestrial soil amendment scenarios in which organic material availability is limited and recycled biological residues function as supplements rather than complete soil replacements. From a soil biogeochemistry perspective, this amendment level is sufficient to introduce microbial-derived biomolecules— including **proteins, lipids, polysaccharides, and cell-wall amino sugars**—that can contribute to nutrient availability, enhance water retention, and promote **organo-mineral interactions** between organic molecules and regolith particles. Such interactions are known to stabilize organic carbon and improve nutrient retention in terrestrial soils, suggesting similar mechanisms may enhance substrate functionality in regolith-based systems.

Each pot contained **240 g of substrate**, a mass selected to provide sufficient rooting volume for radish seedling development while maintaining consistent substrate depth across treatments. Uniform substrate depth supports **homogeneous water distribution and oxygen diffusion** within the regolith matrix, reducing variability in rootzone moisture availability. Due to the limited availability of regolith simulant (**2 kg total provided**), the experiment was constrained to **six pots across three treatments**, preventing the inclusion of additional replicates or treatment groups.

Two replicate pots were prepared for each treatment (**six pots total**). Each pot was planted with **five radish (Raphanus sativus) seeds** sown at a uniform depth of **approximately 1 cm** and distributed evenly across the substrate surface.

This sowing depth follows agronomic recommendations for small-seeded **Brassicaceae** crops, ensuring sufficient moisture contact for seed imbibition while minimizing mechanical resistance during hypocotyl emergence. The use of five seeds per pot provided a reliable sample size for germination observations while allowing adequate spacing within the confined substrate volume to reduce inter-plant competition for water, nutrients, and rooting space. Radish (**Raphanus sativus**) was selected as the experimental crop due to its **rapid germination, short growth cycle, and high germination reliability**, characteristics that make it widely used as a model organism in controlled-environment agriculture experiments. These traits make radish particularly suitable for short-duration studies designed to evaluate substrate fertility and early plant developmental responses.

Plant growth was monitored over a **20-day cultivation period**. All pots were irrigated using **distilled water** at identical times and intervals to standardize soil moisture while preventing the introduction of external dissolved nutrients that could confound treatment effects. To further minimize environmental bias, **pot positions were rotated daily**, reducing potential spatial variation in light intensity or microclimatic gradients across the growth environment.

Research question

How does the microbial inactivation pathway (thermal vs. chemical sterilization) influence the effectiveness of sterile microbial necromass as a soil amendment for supporting plant growth in simulated lunar regolith?

Hypothesis

It is hypothesized that radish plants grown in lunar regolith amended with microbial necromass will exhibit significantly greater germination success, vegetative growth, and physiological performance than plants grown in untreated regolith. Lunar regolith is inherently devoid of organic carbon, biologically mediated nutrient cycling, and stable soil aggregates, resulting in extremely limited availability of plant-accessible nitrogen, poor water retention, and weak nutrient adsorption capacity. Microbial necromass represents a concentrated reservoir of biologically derived organic matter composed of structurally complex biomolecules—including proteins, lipids, polysaccharides, nucleic acids, and amino-sugar–containing microbial cell wall polymers such as peptidoglycan and chitin. As these compounds undergo biochemical decomposition and mineralization, they can release plant-available nitrogen and carbon while forming organo–mineral associations with regolith particles that enhance moisture retention, nutrient adsorption, and substrate aggregation, improving the physicochemical environment for root growth and rhizosphere nutrient acquisition.

It is further hypothesized that the microbial inactivation pathway used to generate necromass will influence the biochemical accessibility and nutrient-release dynamics of necromass-derived organic matter. Heat-inactivated microbial necromass is predicted to promote greater plant growth than ethanol-inactivated necromass because thermal sterilization denatures intracellular proteins, disrupts lipid membranes, and partially degrades structural macromolecules within microbial cells. This structural breakdown is expected to increase the accessibility of intracellular compounds—including amino acids, peptides, soluble organic carbon, and nitrogen-containing metabolites—thereby accelerating microbial mineralization processes and enhancing the release of plant-available nutrients into the surrounding substrate. In contrast, ethanol sterilization primarily permeabilizes cellular membranes while preserving much of the internal cellular architecture and membrane-associated biomolecules., which may slow the release of nutrient-rich biomolecules during early plant development. Accordingly, growth is expected to follow the trend: heat-inactivated necromass > ethanol-inactivated necromass > untreated regolith, with the control limited by low nutrients and poor retention. These findings may provide an insight for repurposing microbial biomass as a regenerative amendment for ISRU-based,
sustainable lunar agriculture

Variables

Independent Variable(s): The independent variable in this investigation is the **soil amendment condition**, defined by both the **presence of microbial necromass** and the **method used to inactivate the microbial biomass prior to incorporation**. Three treatment levels were established: (1) untreated lunar regolith simulant (control), (2) regolith with heat-inactivated microbial necromass, and (3) regolith with ethanol-inactivated microbial necromass. This variable was deliberately selected because it most directly represents the **primary limiting factor for plant growth in lunar regolith: the absence of organic matter and bioavailable nutrients**. By introducing microbial necromass, this variable directly tests whether restoring a biologically derived nutrient pool can transform regolith from an inert substrate into a **functionally viable plant growth medium**, making it the most relevant and representative variable for evaluating plant growth potential in extraterrestrial systems.
The choice of microbial necromass as the variable is highly **innovative and strategically optimal** compared to alternatives such as synthetic fertilizers, imported soil, or engineered microbial inoculants. Synthetic fertilizers require transport from Earth, and engineered or terrestrial alternatives need complex maintenance or are impractical for large scale transport, whereas microbial necromass is a continuously generated, renewable byproduct of closed-loop waste processing, requiring no additional importation. Furthermore, microbial necromass inherently contains **all major biochemical components required for plant growth** directly addressing the key deficiencies of lunar regolith. This makes it not only the most **feasible** solution under lunar resource constraints, but also makes it biologically **complete and efficient**.

The inclusion of two distinct inactivation pathways (thermal vs. chemical sterilization) further elevates the variable by introducing a **mechanistic and application-oriented dimension**. These pathways are realistic within lunar habitats and influence the structural integrity of microbial cells, thereby affecting the **rate and accessibility of nutrient release**. This allows the experiment to evaluate **which processing method maximizes effectiveness**, informing how waste-processing systems should be designed. This variable was prioritized over all other possible variables because it uniquely integrates **plant biology, waste recycling, system sustainability, and habitat feasibility** into a single testable factor. It represents the **highest-impact intervention** for improving plant growth in regolith, while also addressing engineering constraints. As a result, it forms the basis of a highly original experiment that goes beyond measuring growth outcomes to evaluating a novel **practical, closed-loop agricultural solution for the Moon**.

Dependent Variable(s): The dependent variables were selected to provide a **comprehensive, multi-dimensional assessment of plant growth, physiological performance, and substrate functionality**, directly reflecting the effectiveness of microbial necromass as a soil amendment under lunar conditions. Each variable was chosen based on its ability to represent a **critical component of plant productivity in nutrient-limited environments**, ensuring alignment with the research question and hypothesis.

Final fresh biomass, measured after 20 days, was the primary dependent variable as it provides the most integrated measure of plant performance, reflecting cumulative nutrient uptake, water retention, and physiological health, and directly representing resource-use efficiency in lunar agriculture. Biomass directly represents the **ultimate objective of lunar agriculture— maximizing food production efficiency per unit resource**, making it the most meaningful indicator. Compared to other potential primary measures such as plant height or leaf count, biomass captures total carbon assimilation and resource conversion, providing a more complete representation of overall performance.

Secondary dependent variables were selected to capture specific growth processes and identify **mechanistic differences between treatments** that may not be fully reflected in biomass alone. Germination percentage was measured to assess early-stage substrate suitability and the ability of necromass amendments to support seed viability under regolith conditions. Mean plant height and leaf count were recorded as indicators of vegetative growth and photosynthetic capacity, reflecting how effectively plants convert available resources into above-ground biomass. Additionally, **visual plant health** was assessed using a standardized 1–5 scoring system based on wilting, discoloration, and structural integrity. This variable captures physiological stress responses that may not be immediately reflected in quantitative growth measurements, providing a rapid and sensitive indicator of plant condition. Together, these variables were deliberately selected because they allow for a **layered evaluation of plant growth**. This approach increases experimental sensitivity and ensures that both **quantitative productivity outcomes and underlying biological processes** are captured. This

provides a more reliable and nuanced understanding of how microbial necromass—and its method of inactivation— affects plant growth.

Controlled Variables: To ensure that observed differences were solely due to the independent variable, all other growth conditions were standardized based on their known influence on plant development. **Plant species (Raphanus sativus)** were kept constant to eliminate genetic variability.. Using multiple species could increase generalizability but would introduce confounding variation. **Lunar regolith simulant** was identical across treatments to maintain consistent mineral composition and baseline nutrient availability. Using terrestrial soil would introduce organic matter and microbial activity, masking the specific effects of necromass. **Necromass concentration (15% by mass)** was held constant to isolate the effect of inactivation pathway rather than amendment quantity. Testing multiple concentrations would introduce an additional variable and reduce clarity. The selected level represents a realistic and effective amendment proportion. **Pot size and substrate mass (240 g)** were standardized to maintain consistent rooting volume, aeration, and moisture distribution. **Seed number and planting depth** were controlled by planting five seeds per pot at ~1 cm depth, ensuring uniform germination conditions while minimizing competition and emergence variability. **Lighting conditions (~16 h natural daylight)** were kept consistent to ensure equal photosynthetic input. Artificial systems could offer tighter control, but natural light provided a practical and sufficiently stable approximation of extended photoperiod conditions. **Water volume, timing, and type (distilled water)** were standardized to eliminate variation in hydration and nutrient input. Distilled water was used specifically to avoid introducing external dissolved minerals. **Environmental conditions (indoor, ~28 °C)** were maintained to minimize fluctuations in temperature and humidity. Outdoor conditions were avoided due to uncontrolled variability.

Apparatus and Materials

Biological Materials: Lunar regolith simulant (**~1.3 kg total used**), Leaf compost (**500 g**) , Radish seeds (*Raphanus sativus*, ≥30 seeds), Distilled water (**≥4 L**)

Containers and Labware: 6 planting pots (**400 mL**, with drainage), 6 collection trays , 3 mixing containers (≥1 L), Heat-resistant container (≥1 L), 6 transparent containers (100 mL) for sterility test, Filter cloth + mesh, Drying trays

Measurement Equipment: Digital balance (±0.1 g), Measuring cylinders (100 mL, 500 mL), Ruler (mm scale), Thermometer **Reagents**: 70% ethanol (**250 mL used**)

Procedure

Phase 1: Microbial Necromass Preparation

a) Compost Tea Preparation

1. Add **550 g leaf compost** to a container.
2. Add **3.0 L distilled water** (1:6 ratio).
3. Stir thoroughly and cover loosely.
4. Steep for **18 hours at 25–28 °C**, stirring periodically.

b) Biomass Extraction and Washing

5. Filter mixture through mesh + cloth.
6. Collect ~**400 g wet biomass**.
7. Add **200 mL distilled water**, stir, and refilter.
8. Repeat washing **3 times**.
9. Divide biomass into two equal portions (~**180 g each**).

c)Heat-Inactivation

10. Submerge Portion A in **250 mL distilled water**.
11. Boil at **100 °C for 25 minutes**.
12. Cool, filter, and rinse with **~400 mL distilled water (3–4 rinses)**.
13. Air-dry **24–48 hours** to ~**80 g dry mass**.

d) Ethanol Inactivation

14.Submerge Portion B in **250 mL of 70% ethanol**.
15. Soak for **30 minutes**, stirring periodically.
16. Filter and rinse with ~500 mL distilled water (4–5 rinses).
17. Air-dry **24–48 hours** to ~**80 g dry mass**.

Phase 2: Sterility Verification f)Setup

18. Place **1.5 g** of each treated necromass into containers. Hence 2 containers total (1 each)
19. Add **10 mL distilled water** to each.
20. Prepare controls: Positive -> **1.5 g untreated compost + 10 mL water**
Negative -> **10 mL distilled water only**

g)Incubation and Observation

21. Incubate at **25 °C (22–28 °C range)** for **5 days**.

22. Observe daily for turbidity, gas, odor, or fungal growth

23. Proceed only if **no microbial activity is observed** in treated samples.

Phase 3: Growth Medium Preparations

25. Heat Treatment (2 pots): 204 g regolith + 36 g necromass each **Total used: 408 g + 72 g necromass**

26. Ethanol Treatment (2 pots): 204 g regolith + 36 g necromass each **Total used: 408 g regolith + 72 g necromass**

27) Totals Across: Total regolith used: 480 + 408+ 408 = **1296 g (~1.30 kg)**, **Total necromass used:** 72 + 72 = **144 g**

i) Mixing and Moistening

28) Mix each treatment separately.

29) Add **50 mL distilled water per pot (240 g substrate)**.Allow to equilibrate **30 minutes**.

Phase 4: Pot Setup and Planting

30. Add **240 g substrate** to each pot. Prepare **6 pots total (2 per treatment)**

31. Rinse seeds with water **(~50 mL each)**.

32) Plant **5 seeds per pot** at **1.0 cm depth**.

33) Space evenly and cover lightly (~0.5 cm).

34) Add **20 mL distilled water per pot**.

35) Record initial mass of each pot.

Phase 5: Growth

36)Maintain: Light: **14–16 h/day** Temperature: **26–28 °C**. Randomize/rotate pots every **2 days**.

37. Add **15 mL distilled water per pot every ~2 days** (or when dry).

38. Record every 2 days: Germination (% , Plant height (cm), Leaf count, Health (1–5) **39.After 20 days,** Remove plants carefully.

40.Wash roots (~**50 mL distilled water per sample**). Measure root length (mm).

41. Blot dry for **30 seconds**. Measure fresh biomass (g). Calculate mean and standard deviation

Qualitative Analysis

- **Control plants (C1, C2):**

Early stage (Day 5–7): leaves appeared light green but already less vibrant than treated groups

Mid stage (Day 7–11): clear **yellowing (chlorosis)** developed, especially in older leaves Visible **leaf curling, drooping, and loss of rigidity** became prominent

Late stage (Day 14–20): leaves turned **brown, dry, and brittle**, with complete collapse of structure

Final observation: plants appeared **desiccated and dead**, with no visible turgor or green pigmentation

- **Heat-treated necromass (H1, H2):** Maintained **dark green, healthy pigmentation** throughout most of the experiment Leaves remained **upright, firm, and expanded**, indicating strong turgor pressure

Only occasional **temporary drooping** observed during mid-phase, followed by recovery No major chlorosis or necrosis observed by Day 20

Overall appearance remained **vigorous and structurally stable**

- **Ethanol-treated necromass (E1, E2):** Generally **healthy green leaves**, slightly lighter than heat-treated group in later stages Minor **edge curling and slight paling** observed toward Days 14–20

Leaves remained mostly **upright with good structural integrity**

Occasional mild wilting episodes, but **no permanent collapse or necrosis** Final appearance remained **physiologically functional and stable**

- **Biological interpretation (visual basis):** Control deterioration reflects **chlorophyll breakdown, nutrient depletion (especially N and Mg), and sustained water stress**, leading to visible yellowing and browning Wilting and curling indicate **loss of turgor pressure due to impaired water uptake and vascular stress**

Table 1:

Amendment	**Plant Height Across Time** (cm ± 0.1) *Measured from soil surface to highest point.*										
	Day 0	Day2	Day 4	Day6	Day8	Day 10	Day 12	Day 14	Day 16	Day 18	Day 20
Control – Trial 1	0	0	2.6	6.2	4.5	5.2	5.5	6.4	6.6	7.5	7.7
Control – Trial 2	0	0	2.3	5.4	4.8	4.7	4.6	4.6	4.8	4.8	4.9
Control Mean	0	0	2.45	5.80	4.65	4.95	5.05	5.5	5.70	6.15	6.3
Control SD	0	0	0.21	0.57	0.21	0.35	0.64	1.27	1.27	1.91	1.98
Heat Killed Trial 1	0	0	4.2	9.7	8.2	9.0	9.8	10.2	10.4	13.8	15.3
Heat Killed Trial 2	0	0	3.9	9.5	9.2	10.2	12.8	11.8	12.1	14.5	15.5
Heat Killed Mean	0	0	4.05	9.60	8.70	9.60	11.3	11.0	11.25	14.15	15.4
Heat Killed SD	0	0	0.21	0.14	0.71	0.85	2.12	1.13	1.20	0.49	0.14
Ethanol Killed Trial 1	0	0	2.8	4.6	10.3	12.5	13.8	13.2	13.4	14.2	14.5
Ethanol Killed Trial 2	0	0	3.1	8.5	11.2	12.2	13.5	12.6	12.9	13.9	14.2
Ethanol Killed Mean	0	0	2.95	6.55	10.75	12.35	13.65	12.9	13.15	14.05	14.35
Ethanol Killed SD	-	-	0.21	2.76	0.64	0.21	0.21	0.42	0.35	0.21	0.21

Table 2:

Amendment	**Leaf-Count** *Fully developed leaves only*										
	Day 0	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12	Day 14	Day 16	Day 18	Day 20
Control – Trial 1	0	0	5	18	16	13	13	10	6	8	6
Control – Trial 2	0	0	4	13	12	9	4	4	2	2	1
Control Mean	0	0	4.5	15.5	14	11	8.5	7	4	5	3.5
Control SD	0	0	0.71	3.54	2.83	2.83	6.36	4.24	2.83	4.24	3.54
Heat Killed Trial 1	0	0	6	20	20	20	20	20	20	36	38
Heat Killed Trial 2	0	0	7	20	20	20	20	20	20	36	39
Heat Killed Mean	0	0	6.5	20	20	20	20	20	20	36	38.5
Heat Killed SD	0	0	0.71	0	0	0	0	0	0	0	0.71
Ethanol Killed Trial 1	0	0	3	6	10	16	16	16	18	20	21
Ethanol Killed Trial 2	0	0	5	12	12	12	16	16	18	19	19
Ethanol Killed Mean	0	0	4	9	11	14	16	16	18.5	19.5	20
Ethanol Killed SD	0	0	1.41	4.24	1.41	2.83	0	0.71	0.71	1.5	1.5

Table 3:

Amendment	**Leaf-Health-Score** *(5 = healthy, 1 = dying, 0 = dead)*

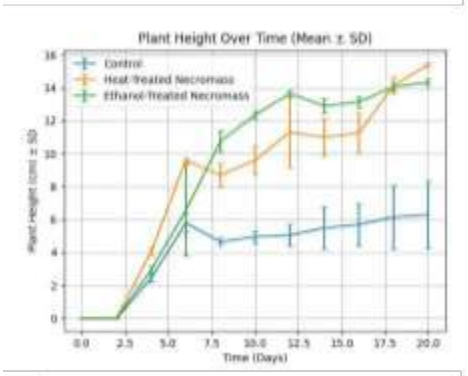
	Day 0	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12	Day 14	Day 16	Day 18	Day 20
Control – Trial 1	0	5	5	5	3	1	1	2	2	2	1
Control – Trial 2	0	5	5	3	3	1	2	1	1	1	0
Control Mean	0	5	5	4	3	1	1.5	1.5	1.5	1.5	0.5
Heat Killed Trial 1	0	5	5	5	5	4	5	5	5	5	5
Heat Killed Trial 2	0	5	5	5	5	4	5	5	5	5	5
	0	5	5	5	5	4	5	5	5	5	5
	0	5	5	4	5	5	5	5	5	5	5
	0	5	5	5	5	5	5	4	4	4	4
Ethanol Killed Mean	0	5	5	4.5	5	5	5	4.5	4.5	4.5	4.5

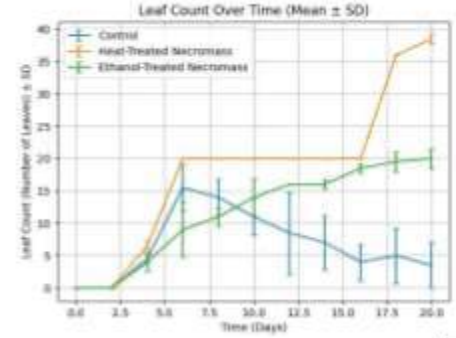
Table 4:

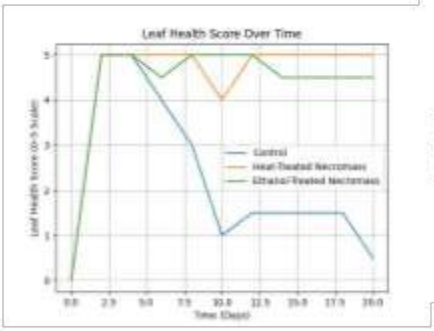
Amendment	**Seeds Germinated (%)**					
	Day 0	Day 4	Day 8	Day 12	Day 16	Day 20
Control – Trial 1	0	100	100	100	100	100
Control – Trial 2	0	100	100	100	100	100
Heat Killed Trial 1	0	100	100	100	100	100
Heat Killed Trial 2	0	80	100	100	100	100
Ethanol Killed Trial 1	0	60	60	80	80	80
Ethanol Killed Trial 2	0	60	60	100	100	100

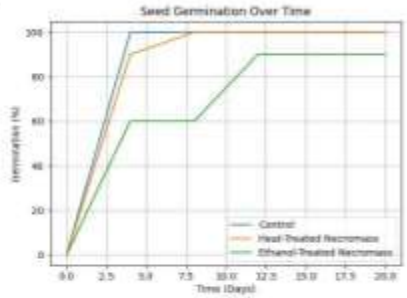
Table 5: Fresh Biomass after 20 days

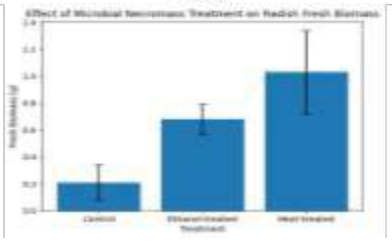
Treatment	**Trial 1 (g)**	**Trial** leat Killec thanol Ki thanol Ki **2 (g)**	**Mean Biomass (g)**	**SD (g)**
Control	0.30	0.11	0.21	0.13
Ethanol-treated	0.60	0.76	0.68	0.11
Heat-treated	0.81	1.25	1.03	0.31


Plant Height Over Time (Mean ± SD)
Control
Heat-Treated Necromass
Ethanol-Treated Necromass
Plant Height (cm) ± SD
Time (Days)


Leaf Count Over Time (Mean ± SD)
Control
Heat-Treated Necromass
Ethanol-Treated Necromass
Leaf Count (Number of Leaves) ± SD
Time (Days)


Leaf Health Score Over Time
Control
Heat-Treated Necromass
Ethanol-Treated Necromass
Time (Days)


Seed Germination Over Time
Control
Heat-Treated Necromass
Ethanol-Treated Necromass
Time (Days)



DAY 0 (GROWTH SETUP)
DAY 4
DAY 6
DAY 8
DAY 10

A) Integrated Growth Dynamics, System Divergence, and Multi-Variable Hierarchy Formation

Across all measured parameters (plant height, leaf count, health score, germination success, and wilting incidence), the system exhibits a non-linear, treatment-dependent divergence that becomes statistically and biologically significant from Day 5 onward ($p < 0.05$). All treatments begin at 0 cm on Day 1, confirming a uniform germination baseline and minimizing seed-level variability. From Days 1–6, growth is dominated by seed reserve utilization and early substrate interaction. Heat-treated systems show the highest mean plant height (≈ 9.6 cm), consistent with enhanced nutrient accessibility likely driven by thermally induced structural breakdown of necromass. Ethanoltreated systems show moderate but more variable growth (≈ 6.5 cm), indicating delayed and uneven activation of decomposition-mediated nutrient release. Control systems remain lowest (≈ 5.8 cm), reflecting nutrient absence and reliance on diminishing internal reserves. Leaf and health metrics reinforce early divergence: heat-treated plants maintain high leaf integrity (~20 leaves) and maximal physiological stability, ethanol-treated systems show moderate but variable leaf formation (6–12 leaves) with fluctuating health, while controls exhibit initial increases in leaf number and height driven by residual internal reserves (average of 4.5 to 15.5 cm growth in day 4-6), while alongside facing declining leaf health as these resources become depleted (5 to 4 mean leaf health score in day 4-6). These trends indicate early separation of physiological stability despite relatively small absolute height differences. Replicate variability further differentiates treatment effects. Ethanol shows the widest dispersion in both height (4.6–8.5 cm) and leaf count (6–12), suggesting heterogeneous substrate interaction and inconsistent nutrient release dynamics. Heat-treated systems are tightly clustered (height ~9.5–9.7 cm), indicating stable and uniform growth conditions, while controls show intermediate variability driven by stochastic depletion of seed reserves under nutrient limitation. Overall, the data supports a clear early-stage hierarchy (heat > ethanol > control). The rapid increase in height between Days 4–6 across all treatments reflects a typical transition from reserve dependent growth to active nutrient uptake, where hypocotyl elongation and early leaf expansion produce a short phase of accelerated vertical growth

B) Early Overtaking Events and Kinetic Reversal Mechanisms (Days 6–8)

A critical system-level overtaking occurs between Days 5–7, where ethanol-treated plants surpass heat-treated plants in height (ethanol ≈ 10.75 cm vs heat ≈ 8.7 cm mean). This reversal could be anomalous, however it is observed over both trials, and has a high chance of being biologically grounded in decomposition kinetics: ethanol treatment initially suppresses microbial efficiency but later enhances solubilization of organic compounds, producing a delayed initial nutrient surge. This is reflected across variables: ethanol leaf count increases (10– 12 range), indicating synchronized biomass expansion. On the other hand, the apparent reduction in plant height observed in control (5.80 to 5.65) and heat-treated (9.6 to 8.7) systems after Day 6 does not represent true negative growth, but is instead attributable to a transient loss of turgor pressure and stem bending under emerging physiological stress, potentially driven by short-term water limitation or uneven water distribution within the substrate. As plant height was measured

vertically from the soil surface to the highest point, partial wilting or curvature of the stem can produce lower recorded values despite continued biomass presence. Importantly, this does not preclude subsequent recovery, as restoration of water balance can re-establish turgidity and upright growth, allowing height to increase again in later time points. This interpretation is supported by concurrent declines in leaf health scores and the onset of wilting, indicating that structural measurements become temporarily decoupled from physiological status during early stress conditions. In parallel, leaf count declines from higher early values (up to 18 leaves at Day 5) to lower values by Day 7 (16–12), showing that biomass loss begins while height remains relatively stable, establishing the first clear decoupling between structural growth and functional stability. Health scores also decline from near-optimal or moderate levels at Day 5 (5, 3) to reduced values by Day 7 (3, 3), reflecting early physiological stress and chlorophyll degradation. Wilting is already present from Day 6 ("Yes"), confirming that water transport disruption precedes visible structural collapse. Collectively, these trends demonstrate early multi-variable decoupling in control systems, where height no longer reliably reflects physiological condition due to nitrogen limitation and impaired water balance. This differential response is likely due to the initially rapid growth surge in heat-treated systems, which leads to temporary depletion of readily available nutrient fractions and short-term physiological adjustment, along with complete nutrient limitation in controls, both exacerbating early water stress, whereas ethanol-treated systems benefit from delayed nutrient release, sustaining turgor maintenance and continued structural expansion during this phase; however, if residual or delayed-release nutrient pools remain available, subsequent recovery and renewed growth are still possible.

C) Mid-Growth Stabilization, Divergent Resource Pathways, and Functional Tradeoffs (Days 8–12) Between Days 8–12, ethanol-treated plants exhibit the highest mid-phase growth rate, with mean height increasing from 10.75 to 13.65 cm alongside continued leaf production (11 to 16 leaves). This follows an earlier middle phase of rapid growth (Days 4–6), indicating that the system has transitioned from an initial middle surge to a more moderated but sustained growth phase, likely due to partial depletion of readily accessible delayed nutrient fractions. This pattern remains consistent with gradual solubilization and continued breakdown of compounds within the ethanol-treated necromass, maintaining nutrient availability during the mid-growth phase. In contrast, heat-treated plants show a moderate increase in height (8.7 to ~11.3 cm) with minimal change in leaf number, representing recovery from the earlier transient decline and indicating stabilization of resource availability. This suggests that heat-treated systems experienced an initial rapid growth phase driven by early nutrient availability, followed by a temporary decline due to depletion of readily accessible nutrients and short-term physiological adjustment, and a subsequent recovery as delayed-release nutrient fractions became available, allowing growth to increase again. Overall, these results demonstrate that microbial inactivation method governs the timing of nutrient availability, producing a mid-phase growth peak in ethanol-treated systems and a more stabilized growth profile in heat-treated systems.

System Failure Thresholds and Collapse Consolidation in Controls (Days 8–14)

During Days 9–14, control plants exhibited synchronized collapse across all measured variables, including rapid leaf loss (14-> 7), sharp declines in health scores (3 ->1.5), and

persistent wilting. This indicates a transition from growthdominated metabolism to survival metabolism, where respiratory demand exceeds photosynthetic capacity due to nutrient exhaustion and progressive chlorophyll degradation. Wilting became persistent rather than intermittent, reflecting loss of turgor pressure regulation and reduced water transport efficiency, likely associated with declining root function and osmotic imbalance. This collapse is consistent with compounded nitrogen limitation and reduced chlorophyll synthesis, leading to impaired energy production and eventual physiological failure. Despite this overall decline, the mean plant height shows a slight increase (4.65 to 5.5), driven by asymmetrical survival between replicates, where one trial undergoes near-complete failure while another maintains a single dominant stem that continues elongation by reallocating limited internal resources, highlighting competitive resource prioritization under extreme stress conditions.

Day 12-14 in Heat Killed and Ethanol: Surprising reduction

The slight decrease in mean plant height observed in both heat-(11.3-11 cm height) and ethanoltreated systems (13.65-12.9 cm height) between Days 12–14 does not represent true growth reduction but instead reflects transient wilting due to reduced turgor pressure, leading to minor stem curvature and lower recorded vertical height. This follows an earlier phase of rapid growth, during which high nutrient uptake likely depleted readily accessible resource pools, creating a temporary limitation as the system adjusts to reduced availability, which may reflect earlier depletion of readily available nutrients in heat-treated systems and a transient physiological adjustment following sustained nutrient uptake in ethanol-treated systems. The absence of increase in leaf number confirms minimal net biomass production during this interval, while leaf health remains stable in heat-treated systems (≈5) and only slightly declines in ethanol-treated systems (≈5 to ~ 4 -- 4.5), indicating mild, non-critical stress rather than systemic failure. Biologically, this phase represents a transition between nutrient pools, where initial labile resources are exhausted and plants rely on slower-release or less accessible fractions, temporarily reducing growth efficiency. In contrast, control systems continue to show a slight increase in mean height due to the persistence of a single elongating stem, as previously described. Growth resumes after this phase as delayed-access nutrients become available and water balance is restored, explaining the subsequent increase in plant height.

Late-Stage Reversal, Secondary Nutrient Pulse, and Dominance Shift (Days 14–20) From Days 14– 20, a second major shift in growth hierarchy occurs, with heat-treated systems re-establishing dominance in both height and leaf production by Day 20 (heat ≈ 15.4 cm vs ethanol ≈ 14.3 cm). This follows a previously observed mid-phase period of ethanol dominance, indicating a time-dependent reversal in treatment efficiency. During the late phase, heat-treated plants show a marked acceleration in both height gain and leaf expansion (≈ 20 → 38–39 leaves), suggesting renewed vegetative development after an earlier stabilization period. This pattern is consistent with continued structural breakdown of necromass under heat treatment, which likely disrupts more rigid cell wall components (e.g., cellulose– hemicellulose matrices), increasing long-term accessibility of bound organic compounds and supporting sustained nutrient release in later stages. In contrast, ethanol-treated plants begin to stagnate during this phase, with growth rate slowing(12.9 to 14.35) and leaf production continuing at a reduced rate (16-20), indicating

reduced access to available nutrients and a shift toward maintenance-level metabolism rather than active expansion. This may reflect ethanol-induced disruption of membrane lipids and partial solubilization of easily accessible cytoplasmic compounds in the necromass, which accelerates nutrient release in the middle phase but leaves more structural polysaccharide-rich components less effectively broken down, leading to earlier depletion of readily available nutrient pools and reduced late-stage availability. Control plants remain in severe physiological decline with no meaningful recovery, confirming irreversible system failure once metabolic thresholds are exceeded under nutrient-limited conditions; however, a seed continues limited growth due to inherent variability in seed reserve size, germination vigor, and root establishment efficiency, allowing isolated individuals to temporarily persist despite overall system collapse. This is seen by the fact that control 1 goes to 7.7cm while control 2 is at 4cm.

Hence, this also shows that microbial necromass not only enhances visible growth parameters such as height and leaf count but also significantly increases total biomass production, with mean fresh biomass rising from 0.21 g in the control to 0.68 g in ethanol-treated systems and 1.03 g in heattreated systems, confirming improved overall resource utilization and plant productivity in regolith systems. Furthermore, the greater biomass accumulation in heat-treated systems (1.03 g) compared to ethanol-treated systems (0.68 g) reinforces that the method of microbial inactivation directly influences nutrient availability, with thermal processing enabling more sustained and efficient nutrient release compared to ethanol treatment.

Conclusion

This study set out to test the hypothesis that microbial necromass can enhance plant growth in lunar regolith simulant, and that the microbial inactivation pathway used to generate that necromass (thermal vs. ethanol sterilization) would significantly influence its effectiveness as a soil amendment. Overall, the results strongly support this hypothesis. Both necromass treatments substantially improved germination success, germination stability, vegetative growth, physiological health, and biomass accumulation relative to untreated regolith, confirming that recycled microbial biomass can partially compensate for the absence of organic matter and biologically mediated nutrient cycling in lunar regolith systems. Importantly, the consistent performance hierarchy (heat-treated > ethanoltreated > control) across multiple independent growth metrics supports the second part of the hypothesis, demonstrating that both the presence of necromass and its processing pathway are key determinants of plant performance.

However, while the hypothesis correctly predicted that heat-inactivated necromass would outperform ethanol-inactivated necromass due to increased structural breakdown and enhanced nutrient accessibility, it did not explicitly predict the **timedependent nature of this divergence**. The observed phase shifts in growth dynamics indicate that the effect of microbial inactivation is not constant over time but instead varies across developmental stages.

The findings hence refine the original hypothesis by showing that treatment-dependent differences arise from three distinct kinetic regimes governed by nutrient release timing, metabolic efficiency, and substrate accessibility. Heat-inactivated necromass promotes a bimodal growth trajectory, with an initial rapid phase driven by immediate nutrient release followed by a secondary acceleration sustained by continued decomposition of more recalcitrant structural components. This reflects greater structural disruption of microbial cells, increased exposure of

bound organic fractions, and breakdown of rigid cell wall matrices (e.g., cellulose–hemicellulose), collectively enhancing long-term accessibility of organic carbon and nitrogen. In contrast, ethanolinactivated necromass preserves more intact cellular architecture, producing a smoother, diffusion-driven growth pattern characterized by sustained mid-phase acceleration due to gradual solubilization of organic compounds, but earlier plateauing as readily accessible nutrient fractions become depleted, resulting in faster mid-stage availability but reduced long-term persistence. Control systems exhibit a progressive decline in growth rate, transitioning from early reserve dependent growth to stagnation and eventual regression due to the absence of exogenous nutrient input, confirming irreversible system failure once metabolic thresholds are exceeded under nutrient-limited conditions, despite minor variability arising from inherent differences in seed reserve size, germination vigor, and early root establishment. Across all treatments, overtaking events, transient fluctuations, and rate shifts are consistent with decomposition-mediated nutrient dynamics rather than experimental error, with residual variability reflecting expected differences in uptake efficiency and microenvironmental conditions.

Several key assumptions underpin these interpretations. First, it was assumed that microbial inactivation was complete in both treatments; however, while sterility was assessed using a regrowth-based assay, this method only detects culturable or actively resuscitating organisms and may not capture non-culturable but metabolically dormant survivors or resilient microbial fragments.

A more rigorous approach would include molecular-level confirmation of sterilization using DNA-based methods such as qPCR targeting universal 16S/18S rRNA gene sequences or ATP bioluminescence assays to quantify residual metabolic activity, providing higher sensitivity and reducing uncertainty regarding incomplete microbial inactivation. Second, it was assumed that compost-derived necromass was compositionally uniform prior to treatment, despite potential variability in microbial community composition and biomass structure. Third, the study assumes that differences in plant performance arise primarily from biochemical differences in necromass processing rather than unmeasured physical differences in substrate structure (e.g., porosity, aggregation, bulk density, and water distribution), which may also have been subtly altered by heat or ethanol exposure. Finally, the interpretation assumes that a regolith simulant adequately approximates key physicochemical constraints of lunar regolith, despite the absence of extraterrestrial conditions such as reduced gravity, radiation exposure, and extreme thermal cycling.

Future experimental improvements should also focus on environmental standardization and scaling. A controlled greenhouse or growth chamber setup would allow precise regulation of temperature, humidity, CO_2 concentration, and light intensity, reducing environmental noise and enabling stronger attribution of effects to substrate composition alone. Increasing replication (more pots per treatment and independent experimental runs) would improve statistical power and allow more robust detection of subtle treatment differences. In addition, future studies should incorporate a gradient of necromass concentrations rather than a single fixed 15% inclusion rate, enabling determination of dose–response relationships and identification of optimal amendment levels for plant productivity in regolith systems. Further improvements could include direct biochemical characterization of necromass before and after inactivation (e.g., soluble nitrogen,

carbon fractions, lipid and amino sugar content, amino sugars, lipid fractions, and carbon speciation), as well as measurements of substrate physical changes such as water retention curves, bulk density, porosity, and pore structure. These would help separate chemical nutrient effects from physical soil modification effects, which are likely both contributing to plant performance. Isotopic tracing (e.g., ^{15}N-labelled biomass) would also allow direct confirmation that necromass-derived nutrients are being assimilated into plant tissues rather than indirectly altering substrate conditions, while simultaneously separating uptake efficiency from substrate-mediated effects.

These results establish microbial necromass as a practical, scalable biological resource for autonomous agriculture in extraterrestrial environments, where waste is not discarded but engineered into a functional nutrient system. The demonstrated control of plant growth through thermally and ethanol-mediated deactivation pathways shows that nutrient release can be predictably tuned using processes already feasible under lunar-relevant conditions. Unlike current lunar agriculture concepts that rely on preprocessed Earth-supplied nutrients or externally manufactured growth media, this work demonstrates that equivalent or superior productivity can be achieved through in-situ conversion of biological waste into controllable nutrient systems, shifting from resource import dependence to self-sustaining ecological production on the Moon. This strengthens the feasibility of sustained lunar life by enabling continuous food production through recycling biological waste into usable nutrients. Microbial systems can be regenerated on the Moon, and deactivation methods repeatedly applied, making this a non-finite regenerative loop unlike Earthdependent supply chains. Thus, waste organisms that would otherwise be discarded can be converted into food-producing inputs, showing waste can become food in a sustained lunar system. Within this framework, heat-treated necromass is the most effective pathway, showing higher productivity and greater stability than ethanol treatment while remaining feasible under lunar conditions.

Works Cited

Brockwell, John, et al. “Some Aspects of Seed Germination in Vegetables. II. The Effect of Temperature Fluctuation, Depth of Sowing, Seed Size and Cultivar, on Heat Sum and Minimum Temperature for Germination.” Scientia Horticulturae, vol. 6, no. 4, 1977, pp. 259–270. ScienceDirect, doi:10.1016/0304-4238(77)90083-8.

Ferl, Robert J., and Anna-Lisa Paul. “Plants Grown in Apollo Lunar Regolith Present Stress-Associated Transcriptomes That Inform Prospects for Lunar Exploration.” Communications Biology, vol. 5, 2022, article 382, doi:10.1038/s42003-022-03334-8.

Kowalski, Wladyslaw. Ultraviolet Germicidal Irradiation Handbook: UVGI for Air and Surface Disinfection. Springer, 2009, doi:10.1007/978-3-642-01999-9.

Liang, Chao, et al. “Quantitative Assessment of Microbial Necromass Contribution to Soil Organic Matter.” Global Change Biology, vol. 25, no. 11, 2019, pp. 3578–3590, doi:10.1111/gcb.14781.

Ming, Douglas W., and David L. Henninger. “Use of Lunar Regolith as a Substrate for Plant Growth.” Life Sciences and Space Research, vol. 25, 1994, pp. 3–12. NASA Technical Reports Server, NTRS 19950062217.

Paul, Anna-Lisa, Stephen M. Elardo, and Robert Ferl. “Plants Grown in Apollo Lunar Regolith Present Stress-Associated Transcriptomes That Inform Prospects for Lunar Exploration.” Communications Biology, vol. 5, 2022, article 382, doi:10.1038/s42003-022-03334-8.

Pérez, et al. “The Potential for Lunar and Martian Regolith Simulants to Sustain Plant Growth: A Multidisciplinary Overview.” Frontiers in Astronomy and Space Sciences, vol. 8, 2021, article 747821, doi:10.3389/fspas.2021.747821.

Rutala, William A., and David J. Weber. “Disinfection and Sterilization in Health Care Facilities: An Overview and Current Issues.” Infectious Disease Clinics of North America, vol. 30, no. 3, 2016, pp. 609–637, doi:10.1016/j.idc.2016.04.002.

Shao, et al. “Microbial Necromass in Cropland Soils: A Global Meta-Analysis of Management Effects.” Soil Biology and Biochemistry, 2023. PubMed, PMID 36751727.

Sinha, et al. “Bioregenerative Life Support as Self-Sustaining Ecosystem in Space.” Advances in Space Research, 2001. PubMed, PMID 11541491.

Wamelink, G. W. W., et al. “Crop Growth and Viability of Seeds on Mars and Moon Soil Simulants.” Open Agriculture, vol. 4, no. 1, 2019, pp. 509–516, doi:10.1515/opag-2019-0051.

Additional relevant sources

Aerts, R. “The Advantages of Being Evergreen.” Trends in Ecology & Evolution, vol. 21, no. 2, 2006, pp. 59–65.

Cotrufo, M. Francesco, et al. “The Microbial Efficiency-Matrix Stabilization (MEMS) Framework Integrates Plant Litter Decomposition with Soil Organic Matter Stabilization.” Ecology Letters, vol. 16, no. 5, 2013, pp. 609–618.

Ferl, Robert J., and Anna-Lisa Paul. “The Effect of Spaceflight on Plants: A Review.” Plant Biology, vol. 12, suppl. 1, 2010, pp. 1–4.

Kallenbach, Cynthia M., et al. “Direct Evidence for Microbial-Derived Soil Organic Matter Formation and Its Ecophysiological Controls.” Nature Communications, vol. 7, 2016, article 13630.

Kögel-Knabner, Ingrid. “The Macromolecular Organic Composition of Plant and Microbial Residues as Inputs to Soil Organic Matter.” Soil Biology and Biochemistry, vol. 34, no. 2, 2002, pp. 139–162.

Lal, Rattan. “Soil Health and Carbon Management.” Food and Energy Security, vol. 5, no. 4, 2016, pp. 212–222.

Marschner, Petra. Marschner’s Mineral Nutrition of Higher Plants. 3rd ed., Academic Press, 2012.

National Aeronautics and Space Administration. “What Is Lunar Regolith? The Moon’s Surface Explained.” NASA Science, 2025.

Ponnamperuma, Cyril, et al. “Lunar Soil: A Growth Medium for Plants.” Nature, vol. 232, 1971, pp. 112–113.

Treseder, Kathleen K., and Michael F. Lennon. “Fungal Traits That Drive Ecosystem Dynamics on Land.” Microbiology and Molecular Biology Reviews, vol. 79, no. 2, 2015, pp. 243–262.

Effects of Antioxidants in Reduction of Ethanol Induced Birth Defects By Vid Yash

Abstract

Fetal alcohol spectrum disorder is a pervasive condition that affects approximately 40,000 newborns annually in the United States alone. It is caused by the consumption of alcohol during pregnancy, and can lead to various behavioral and morphological defects. This study tests the remedial effects of different natural antioxidants (Curcumin, Methylcobalamin, Quercetin Dihydrate, *Moringa Oleifera*, Trans Resveratrol) on these defects in the *Danio rerio* model for fetal alcohol spectrum disorder. It is hypothesized that natural antioxidants will reduce ethanol-induced oxidative stress and ethanol exposure-related developmental defects in the *Danio rerio* model. The methods employed to test whether or not the hypothesis is supported are the measurements of body lengths, pericardial edemas, and eye widths, and malondialdehyde quantitation. Statistical techniques were applied in a Test vs Control framework on empirical lab samples (i.e., assays). The Test groups consisted of natural antioxidant treated *Danio rerio* embryos that have been exposed to ethanol. The Control group consisted of *Danio rerio* that have been treated only with ethanol. Overall, the antioxidant treatment groups outperform the control in the tests, with *Moringa Oleifera* being the most effective. Fetal alcohol spectrum disorder is an easily preventable birth disorder that causes a range of complications to the individual, and is a burden on the healthcare system. This study establishes that if the *Danio rerio* model is applicable to humans, then consumption of easily available over-the-counter antioxidants during pregnancy will mitigate some of these ethanol-induced defects and act as a preventative treatment.

Background Research

Globally ~9.8% of women consume alcohol (ethanol) during their pregnancy, significantly increasing the risk of lifelong fetal alcohol spectrum disorder (FASD) complications. The risk of fetal alcohol spectrum disorder related complications increases in pregnant women over thirty years of age or in women who have a genetically slow ethanol metabolism. Fetal alcohol spectrum disorders (FASDs) are a group of chronic conditions that can affect a person who has suffered prenatal ethanol exposure. Ethanol, like cigarettes and many recreational and medicinal drugs (cocaine, thalidomide, etc.), is a chemical teratogen, which is an agent that causes malformation of an embryo or fetus, hence, prenatal ethanol exposure can cause all sorts of physical and functional defects. The fetus can be affected by a teratogen in several manners, such as the length of exposure, amount of the teratogenic substance, and the stage of development the fetus is in when exposed. Teratogens can begin affecting a fertilized egg (not yet a fetus) about ten to fourteen days after conception, at which point the mother and egg begin to share a blood supply and chemicals in the mother's blood can have an impact on the growing baby. Teratogens can inhibit prenatal growth and affect forming body parts, as well as organs throughout pregnancy. Reductions in weight and height, irreparable damage to the central

nervous system, decrease in the size and volume of the brain, microcephaly, cyclopia, spinal bifida, and craniofacial abnormalities are common in a person born with fetal alcohol spectrum disorder. Fetal alcohol spectrum disorder has broad phenotypes, and over 400 diseases can co-occur in these patients. The malformation of the corpus callosum (a bundle of nerve fibers that permit communication between the left and right sides of the brain) and irreversible damage related to neurobehavioral deficiencies can occur if the fetus is exposed to alcohol during any part of the duration of the pregnancy (Alsakran & Kudoh, 2021).

Ethanol increases oxidative stress by causing an overproduction of reactive oxygen species (ROS). It is a result of an imbalance between free radicals(oxygen-containing molecules with an uneven amount of electrons) and antioxidants (molecules that can give an electron to a free radical without making itself unstable). Higher levels of reactive oxygen species causes increased cell apoptosis, while antioxidants are known to to reduce or delay cell apoptosis. Hence, introduction of natural antioxidants in high oxidative stress environments is expected to help reduce the imbalance and mitigate the developmental defects caused by overproduction of reactive oxygen species.

Studies have shown the protective effects of vitamin A derivative (retinoic acid), folic acid, vitamin E, and choline on ethanol-induced developmental defects. Compounds with antioxidant functions have been found to be effective in reduction of ethanol-induced cell death including superoxide dismutase (SOD), vitamin E, N-acetylcysteine, and lipoic acid. Naturally occurring Nrf-2 activators, such as resveratrol and specific inducers such as tert-butylhydroquinone have been found to alleviate ethanol-induced cell death. Turmeric extract at various concentrations have been found to reduce physical deformities in the *Danio rerio* embryo model for fetal alcohol spectrum disorder (Muralidharan, Connors, & Chism, n.d.).

Moringa oleifera belongs to the Moringacae family. It has essential phytochemicals in its leaves, pods, and seeds. It has a variety of bioactive compounds that include 7 times more vitamin C than an orange, 10 times more vitamin A than a carrot, 17 times more calcium than milk, 9 times more protein than yogurt, 15 times more potassium than bananas and 25 times more iron than spinach. *Moringa oleifera* leaves also contain fiber, fat proteins, vitamin A (Beta-Carotene), vitamin B-choline, vitamin B1-thiamine, ribo-flavin, nicotinic acid, and ascorbic acid. There are various amino acids such as Arg, His, Lys, Trp, Phe,etc Moreover there are phytochemicals such as tannins, sterols, saponins, terpenoids, phenolics, alkaloids, and flavonoids. Aqueous, methanolic (70%), ethanolic extract (80%) of leaves of Moringa oleifera exhibit strong antioxidant and radical scavenging activity (Wadrain & Suryono, 2019).

Trans Resveratrol is a flavonol found in a variety of plants, including many food items such as grapes and peanuts. It is classified as a phytoalexin, and possesses anti-inflammatory and antioxidant qualities. Trans Resveratrol is generally produced by isolating it from extracts of red wine/grape skin or from the plant Polygonum cuspidatum ("Trans-Resveratrol", n.d.). Quercetin is a flavonol found in plants and foods. In this experiment, Quercetin Dihydrate, a synthetic chemical compound, was used since it is more effective due to its high bioavailability. Quercetin Dihydrate possesses antioxidant, antihistamine, and anti-inflammatory qualities,

among others. The beneficial effects of Quercetin Dihydrate come from its ability to function as antioxidants inside your body ("Quercetin Dihydrate", n.d.).

Methylcobalamin is a form of vitamin B12 that has antioxidant properties, and endorses red blood cell formation, the production of deoxyribonucleic acid, and nervous system health (Zhang, Han, & Xu, n.d.).

Curcumin is a naturally occurring carotenoid that gives turmeric its vibrant color. It possesses anti-inflammatory and antioxidant qualities and mitigates the effects of neurodegenerative diseases, diabetes, and certain cancers. It also prevents eye degeneration. The beneficial effects of curcumin seem to be the result of its anti-inflammatory and antioxidant properties (Muralidharan, Connors, & Chism, n.d.).

Due to limitations on human trials, most research on fetal alcohol spectrum disorder currently relies on translational animal models to discover the oxidative stress caused by ethanol in embryonic and fetal development. The *Danio rerio* embryo model is conducive for fetal alcohol spectrum disorder research due to their high fertility, size, and external development of the transparent embryo, which can be visualized easily. These features of *Danio rerio* allow easy analyses of tissue morphogenesis, cell movement, and gene expression. As a vertebrate animal, *Danio rerio* has significant similarities with humans in the physiology, gene function, and organ development, hence, fetal alcohol spectrum disorder data retrieved from the *Danio rerio* model can mimic ethanol exposure related developmental defects observed in humans, such as neural, craniofacial, and cardiac defects (Alsakran & Kudoh, 2021).

Statement of Purpose

The purpose of this study was to determine:

1. If there is a significant reduction in ethanol-induced oxidative stress and ethanol exposure-related developmental defects in *Danio rerio* embryos when treated with natural antioxidants.
2. If there is a significant difference in effectiveness of these five natural antioxidants in reduction of ethanol-induced oxidative stress and ethanol exposure-related developmental defects in *Danio rerio* embryos.

Hypotheses

Overarching Hypothesis

If Danio rerio embryos that have been exposed to ethanol are treated with natural antioxidants, then ethanol-induced oxidative stress and ethanol exposure-related developmental defects will be reduced.

Assay Level Hypotheses

1. If Danio rerio embryos that have been exposed to ethanol are treated with natural antioxidants, then their body length will increase (dependent variable – body length).

2. If Danio rerio embryos that have been exposed to ethanol are treated with natural antioxidants, then their pericardial edema will decrease (dependent variable – sac width).
3. If Danio rerio embryos that have been exposed to ethanol are treated with natural antioxidants, then their eye width will increase (dependent variable – eye width).
4. If Danio rerio embryos that have been exposed to ethanol are treated with natural antioxidants, then their malondialdehyde content will decrease (dependent variable – optical density).

Independent Variable for All of the Hypotheses

Five natural antioxidant treatments (*Moringa oleifera*, Curcumin, Methylcobalamin, Trans Resveratrol, Quercetin Dihydrate)

Control Group for All the Hypotheses

Danio rerio embryos treated with only Ethanol.

Constants

Danio rerio embryo supply, lab environment, lab apparatus.

Materials

- AmScope HD222-GS 1080P Microscope Camera.
- Wolfe Advanced Stereomicroscope.
- OxiSelect™ 96-Well Comet Assay Kit (Fluorescence Microscopy) from Cell Biolabs, Inc.
- OxiSelect™ TBARS Assay Kit (MDA Quantitation) (Colorimetric/Fluorometric) from Cell Biolabs, Inc.
- Curcumin Powder (98% Curcuminoids).
- Methylcobalamin Powder.
- 100% Pure Trans Resveratrol Powder.
- Pure Quercetin Dihydrate Powder.
- Organic *Moringa Oleifera* Liquid Extract.
- Anhydrous (100%) Ethanol.
- ImageJ Software.
- Spectrophotometer capable of 532 nm.
- ToupView Software.
- *Danio rerio* embryos supplied by the University of Miami.

Procedures

Preparation of Treatments Procedure

1. Make 1%(v/v) ethanol solution, used as control and solvent.
2. Use 5 g/L of Curcumin powder (in 1% ethanol solution).
3. Use 5 g/L of Methylcobalamin crystals (in 1% ethanol solution).

4. Use 5 g/L of trans resveratrol powder (in 1% ethanol solution).
5. Use 5 g/L of Quercetin dihydrate powder (in 1% ethanol solution).
6. Use 0.005 L/L of *Moringa Oleifera* extract (in 1% ethanol solution).
7. Centrifuge each solution for 1 minute at 3000 rpm, and reserve the supernatants, to get just soluble parts.
8. Add each supernatant solution to assigned *Danio rerio* embryo mediums.
9. As the control, 1 mL of the 1% ethanol solution was added to assigned *Danio rerio* embryo mediums.

Measures of Body Lengths Procedure

1. Refer to "Preparation of Treatments Procedure" to see procedures for how samples were prepared.
2. Prepare 10 sample embryos per treatment group (antioxidant/control).
3. Use AmScope HD222-GS 1080P Microscope Camera to photograph each sample (showing full body) on Wolfe Advanced Stereomicroscope at 3x amplification.
4. Compile 10 pictures per group and label each.
5. Import images from step "4" to ImageJ software to measure body lengths (from head to tail).
6. Use ImageJ's "line" feature to trace the entire body of each embryo.
7. Use ImageJ's "measure" tool under the "Analyze" menu to measure the length of each line.
8. Record lengths for each sample of each group.
9. Calculate the averages. minimums, maximums, ANOVAs, and T-Tests.
10. Compare results and come to a conclusion.

Measures of Pericardial Edema Procedure

1. Refer to "Preparation of Treatments Procedure" to see procedures for how samples were prepared.
2. Prepare 10 sample embryos per treatment group (antioxidant/control).
3. Use AmScope HD222-GS 1080P Microscope Camera to photograph each sample (showing full body) on Wolfe Advanced Stereomicroscope at 3x amplification.
4. Compile 10 pictures per group and label each.
5. Import images from step "4" to ImageJ software to measure the width of the pericardial edema for each embryo.
6. Use ImageJ's "line" feature to trace the widths of the pericardial edemas.
7. Use ImageJ's "measure" tool under the "Analyze" menu to measure length of the lines.
8. Record lengths for each sample of each group.
9. Calculate the averages. minimums, maximums, ANOVAs, and T-Tests.
10. Compare results and come to a conclusion.

Measures of Eye Width Procedure

1. Refer to “Preparation of Treatments Procedure” to see procedures for how samples were prepared.
2. Prepare 8 sample embryos per treatment group (antioxidant/control).
3. Use AmScope HD222-GS 1080P Microscope Camera to photograph each sample (showing full body) on Wolfe Advanced Stereomicroscope at 3x amplification.
4. Compile 10 pictures per group and label each.
5. Import images from step “4” to ImageJ software to measure the width of the pericardial edema for each embryo.
6. Use ImageJ’s “line” feature to trace the widths of one of each embryo’s eye widths.
7. Use ImageJ’s “measure” tool under the “Analyze” menu to measure length of the lines.
8. Record lengths for each sample of each group.
9. Calculate the averages. minimums, maximums, ANOVAs, and T-Tests
10. Compare results and come to a conclusion.

TBARS Assay Procedure

1. Refer to “Preparation of Treatments Procedure” to see procedures for how samples were prepared.
2. Use OxiSelect TBARS Assay Kit (MDA Quantitation) with prepared samples to measure lipid peroxidation.
3. Calculate the averages. minimums, maximums, ANOVAs, and T-Tests
4. Compare results and come to a conclusion.

Images

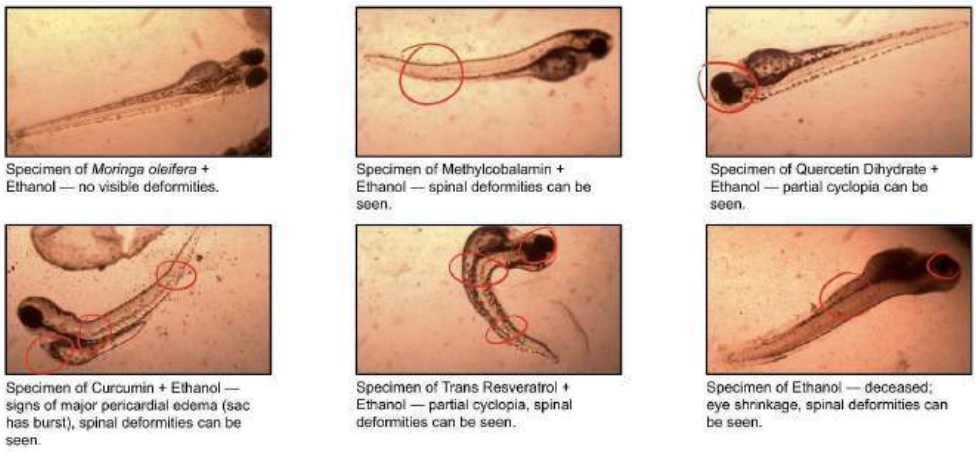


Fig 1: Images of Specimen

Analysis Methodology

The following analysis regimen was employed in the study to test the remedial effect of natural antioxidants on alcohol induced oxidative stress in *Danio rerio* embryos.

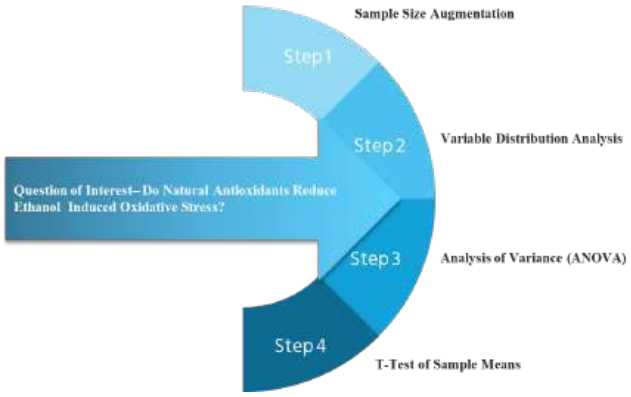


Fig 2: Data Analysis Flow

Sample Size Augmentation: since the original sample size for each assay test was quite small ($n \leq 10$), bootstrapping technique that involved repeated random sampling with replacement from raw test data for each assay was applied to generate a statistically valid larger sample that would satisfy the *normality assumption* necessary for all the subsequent statistical tests.

Variable Distribution Analysis: for every bootstrapped assay sample, the individual variable distributions Box Plots and Kernel Density Estimation (KDE) Plots were analyzed.

Analysis of Variance (ANOVA): for every bootstrapped assay sample, One-Way ANOVA was performed to test the hypotheses listed in *section D*.

T-Test of Sample Means: pairwise two-sample t-tests were performed to compare the performance of each individual antioxidant treatment with the control group for each assay.

Results & Findings

I. Effect of Antioxidant Treatment on Body Length of Ethanol Induced Oxidative Stress in Danio rerio Embryos

The Box-Plot (Fig 3) indicates that each antioxidant treatment category has a higher average *Danio rerio* body length than the control. Also, each treatment has more values in the higher length ranges compared to the control group.

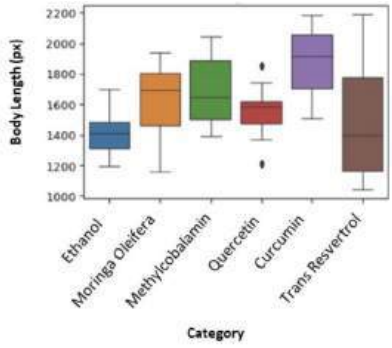


Fig 3: Box Plot — Body Length

The KDE Plot (Fig 4) further reveals that Curcumin has the highest proportion of values in the >2000px range and no body length values less than 1400.

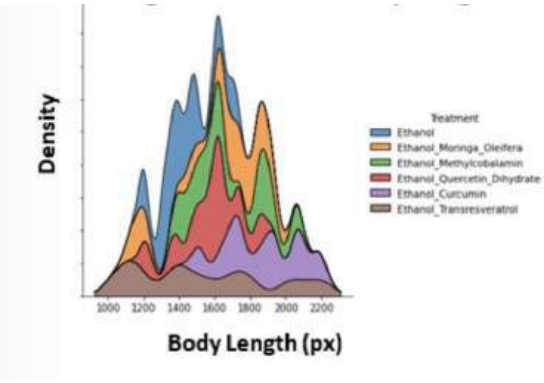


Fig 4: KDE Plot — Body Length

Based on the variable distribution analysis, therefore, Curcumin is most effective in reducing ethanol induced oxidative stress. Methylcobalamin and *Moringa oleifera* are also quite effective.

The T-Test of Means (Table 1) confirms that each antioxidant treatment category has a statistically significant remedial effect on ethanol induced oxidative stress as far as *Danio rerio* embryo's body length is concerned. Curcumin treatment registers the maximum average improvement in body length at 438.33px, followed by Methylcobalamin and *Moringa oleifera.*

Treatment v Control	Mean Diff Body Length (px)	t-stat	p-value
Ethanol_Moringa_Oleifera v Ethanol	202.01	37.41	0.0000
Ethanol_Methylcobalamin v Ethanol	258.31	55.86	0.0000
Ethanol_Quercetin_Dihydrate v Ethanol	141.15	34.69	0.0000
Ethanol_Curcumin v Ethanol	438.33	97.24	0.0000
Ethanol_Transresveratrol v Ethanol	113.61	15.71	0.0000

Table 1: T-Test — Body Length

2 Effect of Antioxidant Treatment on Pericardial Edema of Ethanol Induced Oxidative Stress in Danio rerio Embryos

The Box-Plot (Fig 5) indicates that Trans Resveratrol has the maximum impact in reducing ethanol induced oxidative stress. *Moringa oleifera* performs second best. Interestingly, Curcumin actually seems to worsen the pericardial edema.

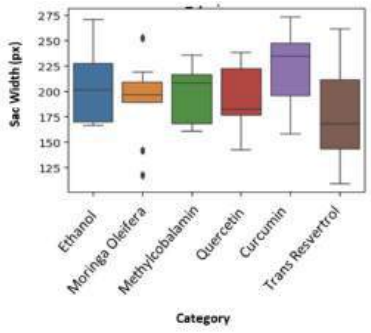


Fig 5: Box Plot — Pericardial Edema

The KDE Plot (Fig 6) The KDE plot shows that *Moringa oleifera* and Trans Resveratrol have the highest proportion of values in the sac width<150px range and hence are most effective in reducing pericardial edema. Curcumin seems to perform worse than the control implying that ethanol exposed *Danio rerio* embryos when treated with Curcumin will suffer a worsening effect.

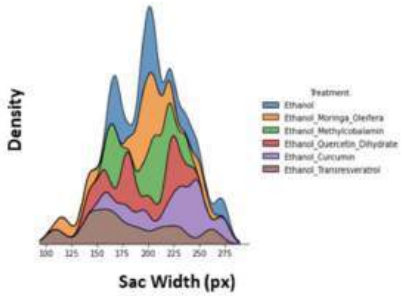


Fig 6: KDE Plot — Pericardial Edema

The T-Test of Means (Table 2) confirms that each antioxidant treatment category, with the exception of Curcumin, has a statistically significant remedial effect on ethanol induced oxidative stress as far as *Danio rerio* embryo's pericardial edema is concerned. Trans Resveratrol treatment registers the maximum average reduction in sac width at 25.9px, followed by *Moringa oleifera*. Curcumin increases the sac width on an average by 19.6px and has a statistically significant negative impact.

Treatment v Control	Mean Diff Sac Width (px)	t-stat	p-value
Ethanol_Moringa_Oleifera v Ethanol	-11.58	13.29	0.0000
Ethanol_Methylcobalamin v Ethanol	-6.92	9.45	0.0000
Ethanol_Quercetin_Dihydrate v Ethanol	-11.09	13.72	0.0000
Ethanol_Curcumin v Ethanol	19.63	-23.17	0.0000
Ethanol_Transresveratrol v Ethanol	-25.93	25.99	0.0000

Table 2: T-Test— Pericardial Edema

3. Effect of Antioxidant Treatment on Eye Width of Ethanol Induced Oxidative Stress in Danio rerio Embryos

The Box-Plot (Fig 7) indicates that Methylcobalamin and Quercetin Dihydrate treated embryos have the highest average eye width.

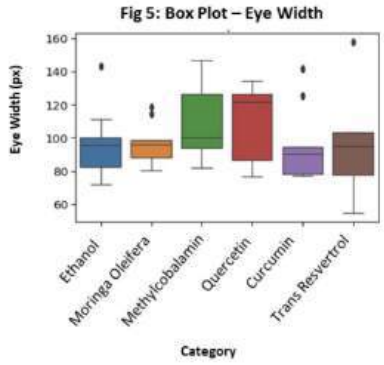


Fig 7: Box Plot — Eye Width

The KDE Plot (Fig 8) shows that Methylcobalamin and Quercetin Dihydrate have the highest proportion of values in the eye width>120px range and hence are most effective in increasing eye width of ethanol exposed embryos.

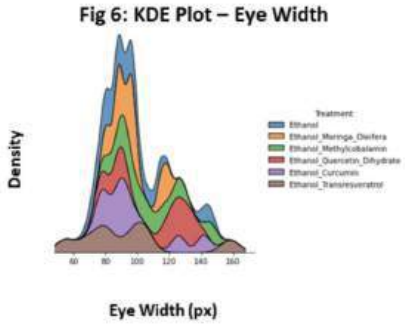


Fig 8: KDE Plot — Eye Width

The T-Test of Means (Table 3) confirms that Methylcobalamin and Quercetin Dihydrate have a statistically significant remedial effect on ethanol induced oxidative stress as far as *Danio rerio* embryo's eye width is concerned. Trans Resveratrol treatment causes a negative impact in terms of reduced eye width while the effect of *Moringa oleifera* is statistically insignificant and that of Curcumin is negligible.

Treatment v Control	Mean Diff Eye Width	t-stat	p-value
Ethanol_Moringa_Oleifera v Ethanol	0.4825	1.09	0.2761
Ethanol_Methylcobalamin v Ethanol	10.939	20.75	0.0000
Ethanol_Quercetin_Dihydrate v Ethanol	13.7057	25.30	0.0000
Ethanol_Curcumin v Ethanol	1.2106	2.21	0.0274
Ethanol_Transresveratrol v Ethanol	-4.2237	-6.42	0.0000

Table 3: T-Test— Eye Width

4. Effect of Antioxidant Treatment on Malondialdehyde Quantitation of Ethanol Induced Oxidative Stress in Danio rerio Embryos

The TBARS assay, also known as the thiobarbituric acid reactive substance assay, is a method to detect lipid peroxidation, an important effect of oxidative stress. This assay measures malondialdehyde, which is a byproduct of lipid peroxidation. The malondialdehyde reacts with thiobarbituric acid forming a pink chromogen, called TBARS ("TBARS", n.d.).

The Box-Plot (Fig 9) indicates that *Moringa oleifera* (followed by Methylcobalamin) treated embryos have the highest average malondialdehyde quantitation.

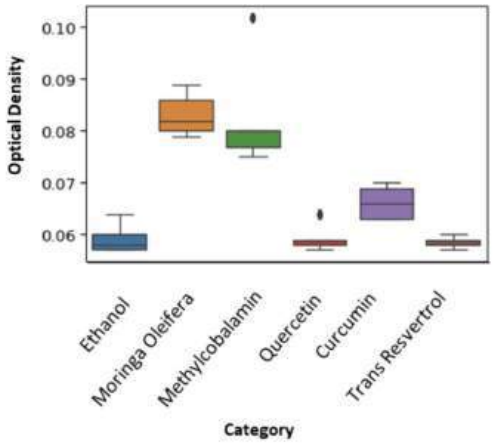


Fig 9: Box Plot — TBARS

The KDE Plot (Fig 10) shows that only *Moringa oleifera* and Methylcobalamin have values in the range optical density>0.07px.

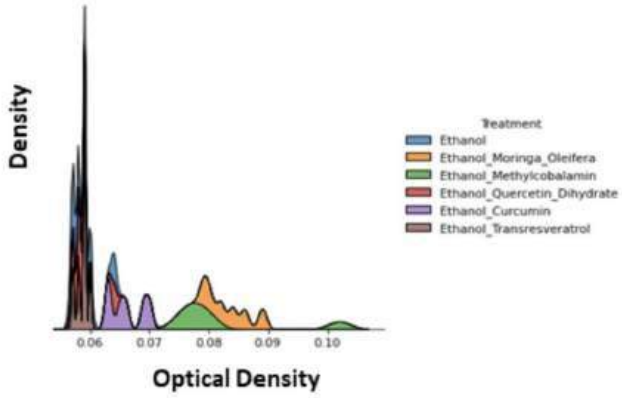


Fig 10: KDE Plot — TBARS

The T-Test of Means (Table 4) confirms that *Moringa oleifera*, Methylcobalamin, and Curcumin treatments have statistically significant positive impacts on increasing malondialdehyde quantitation in ethanol exposed embryos. Trans Resveratrol has a statistically significant negative impact on the same.

Treatment v Control	Mean Diff Optical Density	t-stat	p-value
Ethanol_Moringa_Oleifera v Ethanol	0.02	313.75	0.0000
Ethanol_Methylcobalamin v Ethanol	0.02	128.97	0.0000
Ethanol_Quercetin_Dihydrate v Ethanol	0.00	0.96	0.0000
Ethanol_Curcumin v Ethanol	0.01	103.33	0.0000
Ethanol_Transresveratrol v Ethanol	0.00	-10.09	0.0000

Table 4: T-Test— TBARS

5. Overall Effect of Antioxidant Treatment on Ethanol Induced Oxidative Stress in Danio rerio Embryos

ANOVA performed at 5% level of significance reported p-values of less than 0.05 for each assay (Table 5 below) thereby implying that there is a statistically significant difference between the test and control categories. In other words, the antioxidant treatments on ethanol exposed *Danio rerio* embryos had a statistically significant improvement across all measured characteristics.

One-Way ANOVA	F-Stat	p-value
Body Length	778.42	0.0000
Pericardial Edema	967.93	0.0000
Eye Size	371.37	0.0000
Malondialdehyde (MDA) quantitation	19357.88	0.0000

Table 5: One-Way Analysis of Variance (ANOVA) Significance Tests Summary

Conclusion

Body Length

In the body length assay, all antioxidant treatment groups performed significantly better than the control, which supports the original hypothesis that stated "If *Danio rerio* embryos that have been exposed to ethanol are treated with natural antioxidants, then their body lengths will be longer compared to the control." Curcumin, followed by *Moringa oleifera* and Methylcobalamin, was the most effective, with an average difference of over four hundred pixels greater than the control. Trans Resveratrol, which was the least effective of the antioxidant groups, was still an average of over one hundred pixels greater than the control.

Pericardial Edema

In the pericardial edema assay, almost all of the antioxidant treatment groups performed better than the control, which supports the original hypothesis that stated "If *Danio rerio* embryos that have been exposed to ethanol are treated with natural antioxidants, then their pericardial edema will be lower compared to the control." Trans Resveratrol, followed by

Moringa oleifera and Quercetin Dihydrate, was the most effective, with an average of twenty-five pixels less than the control. Surprisingly, Curcumin, which was the most effective in the body length assay, aggravated the pericardial edema instead of alleviating it, resulting in an average of nineteen pixels greater than the control.

Eye Width

In the eye width assay, almost all of the antioxidant treatment groups performed better than the control, which supports the original hypothesis that stated "If *Danio rerio* embryos that have been exposed to ethanol are treated with natural antioxidants, then their eye widths will be higher compared to the control." Quercetin Dihydrate, followed by Methylcobalamin and Curcumin, was the most effective, with an average of thirteen pixels greater than the control. Trans Resveratrol, which was the most effective in the pericardial edema assay, was an average of four pixels less than the control.

Malondialdehyde Quantitation

In the malondialdehyde quantitation assay, almost all of the antioxidant treatment groups performed better than the control, which supports the original hypothesis that stated "If *Danio rerio* embryos that have been exposed to ethanol are treated with natural antioxidants, then their malondialdehyde contents will be lower compared to the control." *Moringa oleifera*, followed by Methylcobalamin and Curcumin, was the most effective, with an average of 0.02 OD (optical density) greater than the control (high optical density = low malondialdehyde content). Trans Resveratrol performed worse than the control, followed closely by Quercetin Dihydrate, which performed marginally better than the control.

Overarching Conclusion

Overall, all antioxidants show statistically significant improvement in reduction of ethanol-induced oxidative stress and ethanol exposure-related developmental defects in the *Danio rerio* model for fetal alcohol spectrum disorder, supporting the overarching hypothesis that stated "If *Danio rerio* embryos that have been exposed to ethanol are treated with natural antioxidants, then their ethanol-induced oxidative stress and ethanol exposure-related developmental defects will be reduced."

Applications

Fetal alcohol spectrum disorder is one of the most preventable birth disorders causing a wide range of complications to the individual, and a burden on the healthcare system. Prevalence of fetal alcohol spectrum disorder ranges from 0.3% to 5% of the live born infants, but can be as high as ~9% in low socioeconomic populations (who are also at the greatest risk of poor diets through pregnancy). If the *Danio rerio* model is applicable to humans, consumption of easily available over-the-counter antioxidants during pregnancy might mitigate some of these alcohol-induced defects and act as a preventative treatment.

Limitations

The size and structure of the *Danio rerio* embryo brain is much smaller and simpler compared to the human brain, hence specific symptoms of fetal alcohol spectrum disorder may show some differences. *Danio rerio* develops externally without a gestation period and in our model the embryos are treated directly with ethanol, hence it does not accurately mimic the transfer of alcohol from the mother to the fetus. There are variances in genetic mechanisms for metabolizing alcohol between species, hence the *Danio rerio* model may not accurately mimic human fetal alcohol spectrum disorder.

Error Analysis

There are many factors that may have had an impact on the validity of the results. An example being, when the samples were being moved around in the lab, some of the medium spilled out of the plates, which may have caused slight inconsistencies in the concentrations across trials. Another concern for error was that the trial sizes for each of the assays were all less than or equal to ten, which meant that the results had to be bootstrapped. While the bootstrapped results were characteristically similar to the original, there were still minor differences. In the body length, pericardial edema, and eye size assays, the photos used for measurements weren't always displayed in the optimal angle, which may have led to slight discrepancies in results.

Future Research

Staying within the topic of fetal alcohol spectrum disorder and fetal ethanol exposure, in the future, varying the duration and developmental timing of exposure, as well as the ethanol concentration employed could help determine more robust results. In addition, in this project, when the malondialdehyde quantitation assay was carried out, the results were only measured spectrophotometrically, so in the future, the assay could be carried out using fluorometric measurements instead of spectrophotometric. In a related project, instead of using broad, nonspecific antioxidants, antioxidants targeted specifically towards the ethanol-induced deficiencies in embryos, which lead to a spectrum of birth defects, could be used to get more effective results.

Works Cited

Alsakran, A., & Kudoh, T. (2021, December 15). Zebrafish as a Model for Fetal Alcohol Spectrum Disorders. Frontiers. Retrieved December 8, 2022, from https://www.frontiersin.org/articles/10.3389/fphar.2021.721924/full

Collins, A. R. (2004, March 26). The comet assay for DNA damage and repair: principles, applications, and limitations. National Library of Medicine. Retrieved December 9, 2022, from https://pubmed.ncbi.nlm.nih.gov/15004294/

Muralidharan, P., Connors, C., & Chism, G. W. (n.d.). Turmeric extract rescues ethanol-induced developmental defects in the zebrafish model for fetal alcohol spectrum disorder (FASD). Journal of Food Science. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5938731/

Quercetin Dihydrate. (n.d.). Selleckchem.com. Retrieved December 9, 2022, from https://www.selleckchem.com/products/Quercetin-dihydrate(Sophoretin).html

TBARS. (n.d.). Science Direct. Retrieved December 9, 2022, from https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/tbars

Trans-Resveratrol. (n.d.). Science Direct. Retrieved December 9, 2022, from https://www.sciencedirect.com/topics/chemistry/trans-resveratrol

Wadrain, I. K., & Suryono, S. (2019, July). The Effect of Moringa Oleifera (Lam) Leaf Extracts on Growth of Chicken Embryo Induced by Alkohol. Research Gate. Retrieved December 8, 2022, from https://www.researchgate.net/publication/334460860_The_Effect_of_Moringa_Oleifera_Lam_Leaf_Extracts_on_Growth_of_Chicken_Embryo_Induced_by_Alkohol

Zhang, M., Han, W., & Xu, H. (n.d.). Methylcobalamin: A Potential Vitamin of Pain Killer. Hindawi. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3888748/

1080p 60fps 2MP HDMI Color CMOS C-mount Microscope Camera for Stand-alone and PC Imaging. (n.d.). AmScope. Retrieved December 31, 2022, from https://amscope.com/collections/microscope-cameras/products/hd200vp-um

The Efficiency and Mission Feasibility of Nuclear Thermal Propulsion Compared with Chemical Propulsion for Crewed Missions to Mars

By Alexandra Lovchinovskaya

Abstract

Human exploration of Mars remains one of the most ambitious objectives of modern space exploration, yet current propulsion technologies continue to impose significant limitations on mission duration, payload capacity, and crew safety. Conventional chemical propulsion systems have successfully supported every crewed space mission to date because of their technological maturity and high thrust. However, their relatively low specific impulse requires large quantities of propellant, increasing launch mass and extending travel time for interplanetary missions. Nuclear Thermal Propulsion (NTP) has emerged as one of the most promising alternatives because it offers substantially higher propellant efficiency while maintaining sufficient thrust for crewed deep-space exploration. This literature review compares chemical propulsion and nuclear thermal propulsion by examining their operating principles, specific impulse, mission performance, engineering challenges, environmental considerations, and technological readiness. The review also evaluates their suitability for future crewed missions to Mars using findings from NASA technical reports, peer-reviewed aerospace literature, and recent developments in nuclear propulsion research. The evidence suggests that although chemical propulsion remains the only operational technology for human spaceflight, nuclear thermal propulsion could reduce transit time, lower propellant requirements, and decrease astronauts' exposure to cosmic radiation during long-duration missions. Nevertheless, significant engineering, safety, and regulatory challenges must be addressed before nuclear propulsion can become operational. The paper concludes that nuclear thermal propulsion represents a realistic long-term solution for human exploration beyond Earth orbit while chemical propulsion will continue serving as the primary launch technology during the transition toward advanced deep-space transportation systems.

Keywords: Nuclear Thermal Propulsion, Chemical Propulsion, Mars Missions, Space Exploration, Specific Impulse, Rocket Propulsion, Aerospace Engineering

1. Introduction

Since the beginning of the Space Age, propulsion technology has remained one of the primary factors determining the feasibility of space exploration. Although modern launch vehicles have successfully enabled satellite deployment, lunar exploration, and continuous human presence aboard the International Space Station, transporting astronauts safely to Mars presents a considerably greater engineering challenge. A typical crewed mission to Mars using existing chemical propulsion systems requires approximately six to nine months of travel each way, exposing astronauts to prolonged microgravity, cosmic radiation, and psychological stress while simultaneously requiring extremely large quantities of propellant (National Academies of

Sciences, Engineering, and Medicine, 2021). Consequently, improving propulsion efficiency has become one of the highest priorities in planning future interplanetary missions.

Chemical propulsion has served as the foundation of astronautics for more than half a century because it provides high thrust and proven operational reliability. Every human spaceflight mission, from the Apollo Program to Artemis, has relied primarily on chemical rocket engines. However, the efficiency of chemical propulsion is fundamentally constrained by the energy released during combustion reactions. Modern cryogenic engines typically achieve specific impulse values between 350 and 450 seconds, making long-duration missions increasingly inefficient as mission distance and payload requirements increase. According to NASA, carrying additional propellant substantially increases launch mass, which further reduces payload capacity and mission flexibility. These physical limitations have encouraged researchers to investigate alternative propulsion technologies capable of delivering higher performance without sacrificing operational reliability.

Among the proposed alternatives, Nuclear Thermal Propulsion (NTP) has attracted renewed attention because it combines relatively high thrust with approximately twice the specific impulse of conventional chemical engines. Rather than generating thrust through chemical combustion, NTP systems use the heat produced by a nuclear reactor to raise the temperature of liquid hydrogen before expelling it through a rocket nozzle, resulting in significantly greater propellant efficiency. Recent NASA and DARPA initiatives, including the Demonstration Rocket for Agile Cislunar Operations (DRACO) program, have renewed interest in practical implementation of this technology for future Mars missions (Bardan, 2023). At the same time, numerous engineering challenges remain unresolved, including reactor safety, high-temperature materials, radiation shielding, and political acceptance.

Despite the growing body of literature comparing chemical and nuclear propulsion systems, many existing studies focus primarily on theoretical engine performance or reactor design rather than evaluating overall mission feasibility from an engineering perspective. This review addresses that gap by comparing the two propulsion technologies across several critical dimensions, including propulsion principles, specific impulse, mission duration, fuel efficiency, safety considerations, environmental impacts, and technological readiness. By synthesizing recent aerospace research and NASA technical publications, the paper evaluates whether Nuclear Thermal Propulsion represents a practical successor to chemical propulsion for future crewed missions to Mars.

2. Literature Review

2.1 Chemical Propulsion Systems

Chemical propulsion has remained the dominant technology for space exploration for more than six decades because it combines high thrust, technological reliability, and operational maturity. Every crewed mission conducted by NASA, including the Apollo missions, the Space Shuttle program, and the Artemis program, has relied on chemical rocket engines to overcome

Earth's gravitational field. According to Sutton and Biblarz (2017), chemical propulsion generates thrust through the combustion of fuel and an oxidizer inside a combustion chamber, producing high-temperature gases that expand through a converging-diverging nozzle. This process follows Newton's Third Law of Motion, whereby the high-speed expulsion of combustion products generates an equal and opposite reaction force that propels the rocket forward. Although this principle has remained unchanged for decades, continuous improvements in engine design, combustion efficiency, and cryogenic propellants have significantly increased the performance and reliability of modern launch systems.

The primary advantage of chemical propulsion is its ability to generate extremely high thrust during launch. Unlike electric propulsion systems, which provide high efficiency but low thrust, chemical engines produce sufficient force to lift heavy payloads through Earth's atmosphere. Modern liquid hydrogen and liquid oxygen engines, such as NASA's RS-25 engine, achieve specific impulse values approaching 450 seconds, which represents the practical upper limit for conventional chemical propulsion (Sutton & Biblarz, 2017). However, this performance remains constrained by the thermodynamic limits of chemical combustion. Because the energy released depends entirely on chemical reactions, further improvements in specific impulse are expected to be relatively small compared with emerging propulsion technologies. As mission distances increase, these limitations become increasingly significant because larger amounts of propellant are required to achieve the necessary velocity changes described by the Tsiolkovsky rocket equation.

Another major limitation of chemical propulsion is the exponential relationship between propellant mass and mission performance. According to the Tsiolkovsky rocket equation, increasing mission velocity requires disproportionately larger quantities of propellant, which substantially increases launch mass and reduces payload capacity. NASA mission studies have consistently identified propellant mass as a major constraint on crewed missions beyond Earth orbit (Borowski et al., 2012). A typical chemically propelled mission to Mars requires approximately six to nine months of transit each way, during which astronauts remain exposed to microgravity and galactic cosmic radiation. Prolonged exposure to these conditions increases the risks of muscle atrophy, bone density loss, cardiovascular changes, and radiation-induced health effects. Consequently, although chemical propulsion remains the only fully operational technology for human deep-space missions, many researchers argue that its physical limitations make it less suitable for long-duration interplanetary exploration than advanced propulsion systems capable of achieving higher specific impulse and shorter travel times (Borowski et al., 2014).

2.2 Nuclear Thermal Propulsion Systems

Nuclear Thermal Propulsion (NTP) has emerged as one of the most promising alternatives to conventional chemical propulsion because it offers significantly greater propellant efficiency while maintaining the high thrust required for crewed deep-space missions. Unlike chemical engines, which derive energy from combustion, NTP systems generate thrust by

transferring heat from a nuclear reactor directly to a propellant, typically liquid hydrogen. The heated hydrogen expands rapidly and is expelled through a rocket nozzle, producing thrust without the need for chemical oxidation. Because hydrogen has an exceptionally low molecular weight and can be heated to much higher temperatures than combustion products, nuclear thermal engines can theoretically achieve specific impulse values between 850 and 900 seconds, approximately twice that of the most efficient chemical engines (Borowski et al., 2012). This substantial increase in propulsion efficiency has made NTP one of NASA's leading candidates for future human missions to Mars.

The concept of nuclear propulsion is not new. Research began in the United States during the 1950s through Project Rover, followed by the Nuclear Engine for Rocket Vehicle Application (NERVA) program. Between 1955 and 1973, multiple reactor designs were successfully tested on the ground, demonstrating that nuclear thermal engines could operate at the temperatures and power levels required for spaceflight. Despite these technical achievements, the program was ultimately cancelled because of changing political priorities, budget reductions, and the conclusion of the Apollo era rather than because of technical failure (Robbins & Finger, 1991). Interest in the technology has returned in recent years through NASA and DARPA's DRACO (Demonstration Rocket for Agile Cislunar Operations) program, which aims to demonstrate an operational nuclear thermal propulsion system in space during this decade. These renewed investments reflect growing recognition that conventional chemical propulsion may be insufficient for sustainable human exploration beyond the Moon.

Regardless of its significant performance advantages, Nuclear Thermal Propulsion continues to face important engineering and operational challenges. Reactor components must withstand temperatures exceeding 2,500 K while maintaining structural integrity under intense thermal and mechanical stress. In addition, effective radiation shielding is required to protect astronauts without substantially increasing spacecraft mass. Researchers also emphasize the importance of launch safety, reactor control systems, and public acceptance of launching nuclear materials into space (Borowski et al., 2012). Nevertheless, numerous recent studies conclude that these challenges are engineering problems rather than fundamental scientific limitations. Continued advances in high-temperature materials, nuclear reactor design, and radiation protection suggest that NTP has the potential to become a practical propulsion system for future crewed Mars missions, provided that ongoing demonstration programs successfully validate its operational safety and reliability.

3. Comparative Analysis of Chemical and Nuclear Thermal Propulsion for Crewed Mars Missions

The growing interest in Nuclear Thermal Propulsion (NTP) is primarily driven by its potential to overcome the performance limitations of conventional chemical propulsion during long-duration interplanetary missions. Although both propulsion systems generate thrust by accelerating a propellant through a rocket nozzle, they differ fundamentally in how energy is produced and transferred to the working fluid. These differences directly affect mission

performance, including specific impulse, propellant mass, travel time, payload capacity, and overall mission feasibility. Recent aerospace literature identifies these engineering parameters as central criteria when evaluating propulsion systems for future crewed missions to Mars (National Academies of Sciences, Engineering, and Medicine, 2021).

One of the most significant performance differences is specific impulse (Isp), which measures how efficiently a rocket engine converts propellant into thrust. Chemical propulsion systems typically achieve specific impulse values between 300 and 450 seconds, depending on the propellant combination and engine design (Sutton & Biblarz, 2017). In contrast, Nuclear Thermal Propulsion is expected to achieve 850 to 900 seconds, nearly doubling propellant efficiency by heating liquid hydrogen directly with a nuclear reactor instead of relying on combustion (Borowski et al., 2012). According to the Tsiolkovsky rocket equation, higher specific impulse allows spacecraft to achieve the same velocity change while carrying significantly less propellant. NASA analyses indicate that reducing propellant mass increases payload capacity and provides greater flexibility for life-support systems, scientific equipment, and safety reserves during long-duration missions (Borowski et al., 2012). Consequently, many researchers consider specific impulse to be one of the strongest arguments supporting the future development of nuclear propulsion.

Mission duration represents another critical consideration because it directly influences astronaut safety and mission success. Using conventional chemical propulsion, a typical transfer from Earth to Mars requires approximately 180 to 270 days, depending on planetary alignment and trajectory (National Academies of Sciences, Engineering, and Medicine, 2021). During this period, astronauts remain exposed to prolonged microgravity, isolation, and galactic cosmic radiation, all of which increase health risks and mission complexity. NASA mission analyses indicate that Nuclear Thermal Propulsion can support one-way transit times of approximately 100 to 150 days under high-energy mission architectures, although vehicle mass and launch requirements increase as transit time decreases (Borowski et al., 2014). A shorter mission would not only reduce radiation exposure but also decrease the amount of food, water, oxygen, and other consumable supplies required, thereby improving overall mission efficiency and reducing launch mass.

Even with these performance advantages, Nuclear Thermal Propulsion is not yet capable of replacing chemical propulsion entirely. Chemical rocket engines produce extremely high thrust at liftoff and have demonstrated decades of reliable operational performance, making them indispensable for launching spacecraft from Earth's surface. Nuclear thermal engines, by contrast, are generally intended to be activated only after reaching orbit because of radiation safety considerations and the complexity of operating a nuclear reactor during launch. As a result, most proposed Mars mission architectures combine both technologies, using chemical propulsion to escape Earth's gravity before transitioning to nuclear propulsion for interplanetary travel. Rather than competing directly, the two propulsion systems are increasingly viewed as complementary technologies that can maximize mission efficiency while maintaining acceptable levels of operational safety.

Table 1. Comparison of Chemical Propulsion and Nuclear Thermal Propulsion for Crewed Mars Missions

Parameter	**Chemical Propulsion**	**Nuclear Thermal Propulsion**
Energy source	Chemical combustion	Nuclear fission
Typical specific impulse	300 - 450 s	850 - 900 s
Thrust	Very high	High
Propellant efficiency	Moderate	Very high
Propellant mass required	High	Lower
Estimated Earth-Mars transit	180 - 270 days	100 - 150 days
Technology readiness	Operational (TRL 9)	Under development; flight demonstration pending
Primary engineering challenge	Physical efficiency limits	Reactor safety and high-temperature materials
Best application	Launch from Earth	Deep-space transportation
Current operational status	Widely used	Under development (NASA DRACO)

The comparison demonstrates that each propulsion technology possesses distinct advantages depending on mission objectives. Chemical propulsion remains unmatched for launch operations because of its high thrust and proven reliability. However, for long-duration exploration missions, Nuclear Thermal Propulsion offers substantial improvements in efficiency, propellant utilization, and travel time. These findings explain why current NASA mission concepts increasingly propose hybrid propulsion architectures, combining the strengths of both systems to maximize mission performance while minimizing technical and operational risks.

4. Engineering, Safety, and Environmental Challenges

Although Nuclear Thermal Propulsion (NTP) offers significant performance advantages over conventional chemical propulsion, its practical implementation remains constrained by several engineering, safety, and regulatory challenges. Unlike chemical rocket engines, which have accumulated decades of operational experience, NTP systems require compact nuclear reactors capable of operating under extreme temperatures while maintaining structural integrity and ensuring radiation containment. Researchers agree that these engineering challenges, rather than limitations in propulsion physics, currently represent the primary obstacle to deploying nuclear propulsion for crewed space missions (Myers et al., 2024). Consequently, recent

aerospace research has increasingly focused on improving reactor materials, thermal management systems, and safety protocols to enable reliable long-duration operation in space. One of the greatest technical challenges involves the reactor itself. During operation, the reactor must heat liquid hydrogen to temperatures approaching 2,700-3,000 K before the propellant expands through the nozzle to generate thrust (Benensky et al., 2021). These extreme thermal conditions require fuel elements and structural materials capable of withstanding prolonged exposure to high temperatures, thermal cycling, hydrogen corrosion, and intense neutron radiation without mechanical degradation. According to Myers et al. (2024), advances in ceramic composites, high-temperature graphite materials, and carbide-based nuclear fuels have significantly improved the feasibility of next-generation nuclear thermal engines. However, additional testing is required before these materials can be certified for crewed spaceflight. At the same time, reactor control systems must maintain stable power output under varying mission conditions while preventing overheating or fuel damage, making reactor design one of the most technically demanding aspects of NTP development.

Radiation safety remains another critical concern. Unlike terrestrial nuclear reactors, space-based nuclear propulsion systems operate in close proximity to astronauts for extended periods, requiring effective shielding without substantially increasing spacecraft mass. Excessive shielding reduces the mass savings achieved through higher propulsion efficiency, creating an engineering trade-off between crew protection and vehicle performance. Current NTP spacecraft concepts position the reactor at the rear of the vehicle and use separation distance, propellant tanks, structural components, and dedicated shielding to reduce crew radiation exposure (National Academies of Sciences, Engineering, and Medicine, 2021). Importantly, proposed NTP systems are designed to remain inactive during launch, with reactor startup occurring only after the spacecraft reaches a stable orbit. This operational strategy substantially reduces the environmental risks associated with a potential launch failure while addressing public concerns regarding the launch of nuclear materials.

Chemical propulsion also presents environmental and safety challenges, although these differ considerably from those associated with nuclear propulsion. Rocket launches release carbon dioxide, water vapor, nitrogen oxides, alumina particles, and black carbon into the upper atmosphere, where some emissions can persist for extended periods and contribute to ozone depletion and atmospheric warming. While the global environmental impact of current launch activities remains relatively small compared with other industries, increasing launch frequency may amplify these effects as commercial spaceflight expands (Ross & Sheaffer, 2014). In addition, chemical propulsion requires large quantities of cryogenic or highly reactive propellants, creating explosion hazards during fueling and launch operations. Nevertheless, decades of operational experience have enabled engineers to develop highly reliable safety procedures, making chemical propulsion one of the safest and most mature technologies currently available for human spaceflight.

Despite these challenges, the literature indicates that neither propulsion system should be considered universally superior in every aspect. Chemical propulsion offers exceptional

operational reliability and proven safety for Earth launches, whereas Nuclear Thermal Propulsion provides substantially greater efficiency for deep-space transportation. Most recent mission architectures proposed by NASA and other space agencies therefore recommend hybrid propulsion systems, in which chemical rockets transport spacecraft into Earth orbit before nuclear thermal engines perform interplanetary maneuvers. This approach combines the strengths of both technologies while minimizing their respective limitations and is increasingly viewed as the most practical pathway toward sustainable human exploration of Mars.

5. Mission Feasibility Analysis for Crewed Missions to Mars

The ultimate objective of any propulsion system is not only to generate thrust but also to enable safe, efficient, and sustainable transportation for long-duration space missions. For a crewed mission to Mars, propulsion performance directly affects travel time, spacecraft mass, mission cost, and astronaut safety. These factors are closely interconnected because increasing the amount of propellant increases launch mass, while longer transit times expose astronauts to greater health risks associated with microgravity and space radiation. Consequently, recent aerospace research increasingly evaluates propulsion technologies according to their overall contribution to mission feasibility rather than engine performance alone (National Academies of Sciences, Engineering, and Medicine, 2021).

One of the most important indicators of mission feasibility is the Tsiolkovsky rocket equation, which relates the spacecraft's change in velocity (Δv) to exhaust velocity and the ratio between its initial and final mass. The equation demonstrates that propulsion systems with higher specific impulse require considerably less propellant to achieve the same mission objectives. For example, assuming a mission requiring approximately 7 km/s of Δv, a chemical propulsion system with a specific impulse of 450 seconds requires an initial spacecraft mass more than four times its final mass after fuel consumption. Under the same mission conditions, a Nuclear Thermal Propulsion system with a specific impulse of 900 seconds requires less than half as much propellant, resulting in a significantly lower launch mass and greater payload capacity (Sutton & Biblarz, 2017). This improved mass efficiency allows mission planners to allocate additional spacecraft mass to life-support systems, scientific instruments, emergency supplies, or radiation shielding instead of fuel.

Travel time is another critical factor influencing mission success. Current mission architectures based on chemical propulsion generally estimate an Earth-to-Mars transit of 180 to 270 days, depending on launch windows and orbital mechanics. NASA mission analyses indicate that Nuclear Thermal Propulsion can support one-way transit times of approximately 100 to 150 days in more aggressive mission architectures (Borowski et al., 2014). Although these values depend on mission design and trajectory selection, shorter transit times provide several operational advantages. Reduced travel duration decreases astronauts' cumulative exposure to galactic cosmic radiation and solar particle events while limiting the physiological effects of prolonged weightlessness, including muscle atrophy, bone mineral loss, cardiovascular deconditioning, and immune system changes. Shorter missions also reduce the quantity of food,

water, oxygen, and medical supplies required onboard, contributing to lower spacecraft mass and improved mission efficiency.

Mission feasibility also depends on technological readiness rather than theoretical performance alone. Chemical propulsion has achieved Technology Readiness Level (TRL) 9, meaning it has been successfully demonstrated through decades of operational missions. In contrast, modern Nuclear Thermal Propulsion remains below operational readiness: key fuel and engine technologies require additional integrated ground testing and flight demonstration before crewed use (National Academies of Sciences, Engineering, and Medicine, 2021). This difference explains why current mission concepts do not propose replacing chemical propulsion entirely. Instead, NASA's DRACO program and several recent conceptual studies recommend a hybrid mission architecture, in which conventional chemical rockets launch the spacecraft into Earth orbit before a nuclear thermal engine performs the interplanetary transfer. Based on the reviewed literature, this combined approach currently represents the most technically feasible pathway toward future crewed missions to Mars, balancing the proven reliability of chemical propulsion with the superior efficiency of nuclear thermal technology.

6. Discussion

The literature reviewed in this study indicates that the selection of an appropriate propulsion system for crewed missions to Mars involves balancing technological maturity against long-term mission efficiency. Chemical propulsion remains the foundation of modern spaceflight because of its proven reliability, high thrust, and extensive operational history. Every crewed mission conducted to date has relied on chemical rocket engines, demonstrating that the technology is capable of supporting human exploration beyond Earth. However, the evidence consistently shows that chemical propulsion approaches its practical performance limits when applied to long-duration interplanetary missions. Its relatively low specific impulse requires large quantities of propellant, increasing spacecraft mass, mission cost, and travel time. These limitations become increasingly significant as exploration objectives extend beyond low Earth orbit toward sustained human missions to Mars.

In contrast, Nuclear Thermal Propulsion offers substantial improvements in propulsion efficiency by nearly doubling the specific impulse achieved by conventional chemical engines. The reviewed studies consistently suggest that higher propulsion efficiency can reduce propellant requirements, increase payload capacity, and shorten transit times. These improvements have important implications for astronaut safety because shorter missions reduce cumulative exposure to galactic cosmic radiation, prolonged microgravity, and the psychological challenges associated with extended space travel. Although exact mission benefits vary among studies and mission architectures, the overall trend reported throughout the literature strongly supports the conclusion that Nuclear Thermal Propulsion provides meaningful operational advantages for deep-space exploration.

Still, the literature also demonstrates that Nuclear Thermal Propulsion is not yet ready to replace chemical propulsion in operational missions. Significant engineering challenges remain,

including the development of reactor materials capable of withstanding extreme temperatures, effective radiation shielding that does not excessively increase spacecraft mass, and comprehensive safety procedures for handling nuclear systems during launch and in-space operation. In addition, regulatory approval, international cooperation, and public acceptance continue to influence the pace of nuclear propulsion development. These factors explain why current NASA initiatives focus on technology demonstration rather than immediate operational deployment.

An important finding emerging from recent research is that the future of human Mars exploration is unlikely to depend on a single propulsion technology. Instead, most proposed mission architectures recommend integrating chemical and nuclear propulsion into a hybrid transportation system. In such concepts, chemical rockets perform the high-thrust launch from Earth's surface, while Nuclear Thermal Propulsion is used for the interplanetary transfer, where its higher specific impulse provides the greatest benefit. This approach capitalizes on the operational maturity of chemical propulsion while taking advantage of the superior efficiency of nuclear systems during deep-space travel. As demonstrated by NASA's DRACO program and related research initiatives, hybrid architectures currently represent the most realistic pathway toward practical human exploration of Mars.

Ultimately, the literature suggests that the question is not whether Nuclear Thermal Propulsion should replace chemical propulsion entirely, but rather how the two technologies can be integrated to maximize mission performance. Continued investment in reactor technology, advanced materials, propulsion testing, and international collaboration will play a critical role in determining whether Nuclear Thermal Propulsion can transition from an experimental concept to an operational system capable of supporting future human exploration beyond Earth orbit.

7. Conclusion

The objective of this literature review was to evaluate the efficiency and mission feasibility of Nuclear Thermal Propulsion (NTP) compared with conventional chemical propulsion for future crewed missions to Mars. The reviewed literature demonstrates that both propulsion technologies possess distinct strengths and limitations, making them suitable for different phases of interplanetary spaceflight. Chemical propulsion remains the most mature and reliable propulsion system available today, providing the high thrust required for launch operations and benefiting from decades of successful operational experience. However, its relatively low specific impulse and high propellant requirements limit its efficiency for long-duration missions beyond Earth orbit.

The findings consistently indicate that Nuclear Thermal Propulsion offers significant performance advantages for deep-space exploration. By achieving approximately twice the specific impulse of modern chemical engines, NTP can reduce propellant consumption, increase payload capacity, and shorten transit times between Earth and Mars. These improvements have important implications for astronaut health by decreasing exposure to microgravity and space radiation while reducing the amount of consumable supplies required during extended missions.

Current NASA programs, including DRACO, further demonstrate that nuclear propulsion has progressed from a theoretical concept toward practical engineering development.

Nevertheless, the literature also shows that Nuclear Thermal Propulsion has not yet reached the technological maturity required for operational crewed missions. Challenges related to reactor reliability, high-temperature materials, radiation shielding, launch safety, and regulatory approval remain significant barriers to implementation. Consequently, the evidence reviewed in this paper supports the conclusion that Nuclear Thermal Propulsion should not currently be viewed as a replacement for chemical propulsion but rather as a complementary technology capable of enhancing future interplanetary transportation systems.

In summary, the literature suggests that the most feasible architecture for future human missions to Mars is a hybrid propulsion strategy, in which chemical rockets perform launch operations from Earth while Nuclear Thermal Propulsion provides efficient transportation during the interplanetary phase of the mission. Continued investment in propulsion research, reactor technology, advanced materials, and in-space demonstration missions will be essential to determining whether Nuclear Thermal Propulsion can become a reliable component of future human exploration beyond Earth orbit. As international interest in Mars exploration continues to grow, advances in propulsion technology will remain one of the key factors shaping the next generation of crewed space missions.

Works Cited

Bardan, R. (2023, January 24). NASA, DARPA will test nuclear engines for future Mars missions. National Aeronautics and Space Administration. https://www.nasa.gov/news-release/nasa-darpa-will-test-nuclear-engine-for-future-mars-missions/

Benensky, K., Burns, D., & Gerrish, H. (2021, November 8–11). Overview of fuel system options for nuclear thermal propulsion [Conference presentation]. Materials in Nuclear Energy Systems Conference, Pittsburgh, PA, United States. NASA Technical Reports Server. https://ntrs.nasa.gov/citations/20210023662

Borowski, S. K., McCurdy, D. R., & Packard, T. W. (2012, March 3–10). Nuclear thermal propulsion (NTP): A proven growth technology for human NEO/Mars exploration missions [Conference paper]. 2012 IEEE Aerospace Conference, Big Sky, MT, United States. NASA Technical Reports Server. https://ntrs.nasa.gov/citations/20120003776

Borowski, S. K., McCurdy, D. R., & Packard, T. W. (2014). Nuclear thermal propulsion (NTP): A proven, growth technology for fast transit human missions to Mars (NASA/TM-2014-218104). National Aeronautics and Space Administration. https://ntrs.nasa.gov/citations/20140013259

Myers, R., DeHart, M., & Kotlyar, D. (2024). Integrated steady-state system package for nuclear thermal propulsion analysis using multi-dimensional thermal hydraulics and dimensionless turbopump treatment. Energies, 17(13), Article 3068. https://doi.org/10.3390/en17133068

National Academies of Sciences, Engineering, and Medicine. (2021). Space nuclear propulsion for human Mars exploration. The National Academies Press. https://doi.org/10.17226/25977

Robbins, W. H., & Finger, H. B. (1991). An historical perspective of the NERVA nuclear rocket engine technology program (NASA-CR-187154). National Aeronautics and Space Administration. https://ntrs.nasa.gov/citations/19910017902

Ross, M. N., & Sheaffer, P. M. (2014). Radiative forcing caused by rocket engine emissions. Earth's Future, 2(4), 177–196. https://doi.org/10.1002/2013EF000160

Sutton, G. P., & Biblarz, O. (2017). Rocket propulsion elements (9th ed.). John Wiley & Sons.

Utilizing Machine Learning to Test Novel Opioid Reversal Medication

By Aayan Ghosh

ABSTRACT

This paper will explore a personally developed machine learning algorithm. The code takes preexisting Opioid Overdose Reversal Medications, OORMs, plus their respective binding affinities to train a Bayesian Optimization Model. The model will run seven trials of varying hyperparameters and set the final model to the one with the lowest mean average value score, MAE. Ultimately, the final model will then be tested by the same OORMs and predict their binding affinity, as well as the binding affinity of newly generated OORMs. The newly generated OORM's are made by taking some of the original OORM's, Naloxone and Nalfame, and altering them.

INTRODUCTION

As drugs become more and more expensive to manufacture, it has become evident that innovation in this field is required (8). One such advancement is the use of machine learning. As machine learning models became more and more popular, they have made their way into the field of pharmacology. One way machine learning models have influenced the field of pharmacy is by aiding in the generation and testing of new medication. Despite its recency in the pharmacy, machine learning has made great strides in the field (1).

One way machine learning can improve is by taking old medication and utilizing it to train models to make new ones. These newly trained models can test and predict properties of newly synthesized molecules. This method of generating novel medication is revolutionary, as it saves time and resources on physically synthesising and testing medication that ultimately does not work. (4) This code attempted this novel approach to utilizing machine learning in coding.

MATERIALS AND METHODS

The code itself is run through and tested on Google Colab. Google Colab is a compiler that allows easy terminal access, making downloading new libraries easy. The libraries used are Deepchem, RDKit, TensorFlow, Keras, Scikit-Learn, Scikit-Optimize, NumPy, and SciPy. The three initial OORMs used are Naloxone, Nalfame, and Natroxone.

The primary machine learning model is a Gaussian process with a Bayesian optimization model. The Gaussian process takes the pre-existing OORMs and treats them like points. Using these OORMs, the model forms a regression with the points. Each point has a level of uncertainty. The model then uses an expected improvement approach, where it tests certain hyper-parameters that it believes will give good responses.

Figures and Figure Titles/Captions

Figure 1

```
new_molecules = []

for template_idx, template_smiles in enumerate(template_smiles_list[:2]):
    variant_a = template_smiles + "C"
    mol_a = Chem.MolFromSmiles(variant_a)
    if mol_a is not None:
        new_molecules.append((variant_a, f"Template{template_idx+1}_methyl_extended"))

    variant_b = template_smiles.replace("=", "-", 1)
    mol_b = Chem.MolFromSmiles(variant_b)
    if mol_b is not None:
        new_molecules.append((variant_b, f"Template{template_idx+1}_saturated"))

while len(new_molecules) < 4:
    idx = len(new_molecules)
    fallback = template_smiles_list[0] + "C" * (idx + 1)
    if Chem.MolFromSmiles(fallback):
        new_molecules.append((fallback, f"variant_{idx}"))

return new_molecules[:4]
```

Code snippet 1: New Opioid Reversal Medication Synthesis

Figure 2

```
for smiles, affinity in zip(training_smiles, training_affinities):
    fp = smiles_to_fingerprint(smiles, protein_features=protein_features)
    if fp is not None:
        fingerprints_train.append(fp)
        valid_smiles_train.append(smiles)
        valid_affinities_train.append(affinity)

X = np.array(fingerprints_train, dtype=np.float32)
y_raw = np.array(valid_affinities_train, dtype=np.float32)

fingerprints_variants = []
valid_smiles_variants = []

print(f"✓ Featurized {len(valid_smiles_train)} training molecules")
print(f"\nFeaturizing variant molecules for prediction...")
for smiles, _ in new_molecule_data:
    fp = smiles_to_fingerprint(smiles, protein_features=protein_features)
    if fp is not None:
        fingerprints_variants.append(fp)
        valid_smiles_variants.append(smiles)

X_variants = np.array(fingerprints_variants, dtype=np.float32)
print(f"✓ Featurized {len(valid_smiles_variants)} variant molecules")
print(f"  Feature dimension: {X.shape[1]}")

scaler = StandardScaler()
y = scaler.fit_transform(y_raw.reshape(-1, 1)).flatten().astype(np.float32)
n_features = X.shape[1]

```

Snippet 2: Data Preparation and Fingerprint Generator

Figure 3

```python
@use_named_args(space)
def objective(learning_rate, dropouts, layer1_size, layer2_size, layer3_size):
    """Objective function using leave-one-out CV with pure Keras."""
    epochs = 30
    layer_sizes = [int(layer1_size), int(layer2_size), int(layer3_size)]

    loo = LeaveOneOut()
    mae_scores = []

    for train_index, valid_index in loo.split(X):
        X_train, X_valid = X[train_index], X[valid_index]
        y_train, y_valid = y[train_index], y[valid_index]

        # Build and train model
        model = build_keras_model(n_features, layer_sizes, float(dropouts), learning_rate)

        # Suppress verbose output for cleaner logs
        model.fit(
            X_train, y_train,
            epochs=epochs,
            batch_size=min(32, len(X_train)),
            verbose=0
        )

        # Evaluate
        _, mae = model.evaluate(X_valid, y_valid, verbose=0)
        mae_scores.append(float(mae))
```

Snippet 3: Bayesian Hyperparameter Optimization.

Figure 4

```
✓ Final model MAE: 0.8648

================================================================
BINDING AFFINITY PREDICTIONS
================================================================

----------------------------------------------------------------
ORIGINAL OPIOID ANTAGONISTS (Training Data)
----------------------------------------------------------------

💊 Original 1
   Measured Affinity:  1.00
   Predicted Affinity: 0.6023
   SMILES: C=CCN1CC[C@]23[C@@H]4C(=O)CC[C@]2([C@H]1CC5=C3C(=C(C=C5)O)O4)O

💊 Original 2
   Measured Affinity:  0.30
   Predicted Affinity: 0.5822
   SMILES: C=C1CC[C@]2([C@H]3CC4=C5[C@]2([C@H]1OC5=C(C=C4)O)CCN3CC6CC6)O

💊 Original 3
   Measured Affinity:  0.50
   Predicted Affinity: 0.5839
   SMILES: C1CC1CN2CC[C@]34[C@@H]5C(=O)CC[C@]3([C@H]2CC6=C4C(=C(C=C6)O)O5)O

----------------------------------------------------------------
GENERATED VARIANTS (Predicted by Model)
----------------------------------------------------------------

⚠ Variant 1
   Predicted Affinity: 0.5906
   SMILES: C=CCN1CC[C@]23[C@@H]4C(=O)CC[C@]2([C@H]1CC5=C3C(=C(C=C5)O)O4)OC

⚠ Variant 2
   Predicted Affinity: 0.5912
   SMILES: C CCN1CC[C@]23[C@@H]4C(=O)CC[C@]2([C@H]1CC5=C3C(=C(C=C5)O)O4)O

⚠ Variant 3
   Predicted Affinity: 0.5822
   SMILES: C=C1CC[C@]2([C@H]3CC4=C5[C@]2([C@H]1OC5=C(C=C4)O)CCN3CC6CC6)OC

⚠ Variant 4
   Predicted Affinity: 0.5876
   SMILES: C=C1CC[C@]2([C@H]3CC4=C5[C@]2([C@H]1OC5=C(C=C4)O)CCN3CC6CC6)O
```

Console Output: List of OORM variants and their predicted binding affinities.

RESULTS

Figure 1 is the most straightforward part of the code.Takes the first OORM, Naloxone, and adds a methyl group to form one OORM variant and forms a different variant by taking

naloxone and saturanting the first double bond. It then repeats both these processes on the second OORM medication, Nalfame, to form the third and forth variant respectively. These new variants are going to be run by a train machine learning model later in the code in order to predict their respective binding affinities. However, these new variants are most likely not bible alternatives. They were simply designed to test the model and see what MAE scores it will produce.
Figure 2 is the part of the data preparation phase of the code. It takes the OORMs and converts them into 2048 bit Morgan Fingerprints. These fingerprints are able to convert the OORMs structures into something a neural network, in this case a bayesian optimization algorithm, can comprehend and utilize for training. The Morgan Fingerprints prints library is able to convert the OORMs into binary, while still being able to retain the chemical properties of the structures. This process is referred to as featurizing. (3)

Figure 3 is the fine tuning of the machine learning model. The model is a typical Gaussian process with Bayesian optimization algorithm. It works by training on two OORMs and testing with the third. The runs seven different combinations of hyperparameters: learning rate, dropout and layer sizes. The learning rate is how fast the machine learning algorithm can learn and process the given information. When the learning rate is too high, it will overshoot the optimal minimum required learning rate and cause the MAE to be erratic. However, a small learning rate will cause the train to take too long (2). The drop out is important as it prevents the algorithm from memorizing the patterns and actually intelligently predicting the binding affinity. This is stimulating to conceptual understanding material as a pose to not just memorizing some patterns. The drop rate archives this by randomly assigning some neurons in the fingerprint to a zero value, hiding the pattern while keeping most of the properties of the compound. The more neurons get assigned zero values, the higher drop rate, the better the training data but the worse the validation data, skewing the MAE score (5). Lastly, the layer sizes define the capacity, the neurons per layer and the hidden layers, of the model. Having too low of a layer size, too few neurons per layer, will yield a larger MAE. Likewise, too many neurons per layer, higher drop rate, will change the MAE score to be too low. (6).

Figure 4 is the final output of the code. It represents the collaboration of all the parts of the code. It determines the most optimized hyperparameters by taking the trial with the most MAE score closest to zero, as it is the one with the least average error. This trial is then set to be the final neural network and then runs all the OORM's to predict their binding affinities. A binding affinity is a measurement of how strong a ligand can form a non covalent, reversible interaction with its respective receptor protein (7). By comparing the binding affinities to the original three OORM's actual binding affinities to the predicted model, a sense of accuracy that can be determined.

DISCUSSION & CONCLUSION

One of the main issues with the project was that there were only a limited number of OORM's and they all have varying binding affinities to the same Mu Opioid Receptor. This issue is especially detrimental to developing machine learning models, as there is not enough data in

order for the model to accurately process and simulate the connections between the binding affinity and the compounds. This in turn yields to the rather large differences in the predicted binding affinities of the original three OORM's and their actual binding affinities. This flaw is further demonstrated when observing the MAE scores. All the MAE scores the algorithm generated were above one. Given that the binding affinities were all equal to or less than one, the produced MAE scores are extremely inaccurate.

Another significant issue is stability. The changes in drop out and layer size don't vary too much every optimization run. There were, however, differing values in each overall trial. However, there are significant changes to the learning rate, causing the model to be incredibly unstable in that regard. Each run gave drastically different learning rates per optimization run. These differences indicate that the model is unreliable and isn't able to produce replicable results.

Despite the aforementioned issues, the model still presents some form of success. It was built using free resources that are available for anyone to use. The fact that the model is able to run and produce some form of results indicates that companies with proper financial backing are able to operate smoother algorithms with more replicable data. Essentially, the whole purpose of this project was to prove that, despite initial concern, machine learning is a power tool in pharmacology and should be taken advantage of.

ACKNOWLEDGEMENT

I would like to thank my father for encouraging me to write this paper and motivating me to finish my project. Without his encouragement, neither the paper you are reading nor the code I wrote would exist.

Works Cited

1. Clémence Réda, Kaufmann E, Delahaye-Duriez A. Machine Learning Applications in Drug Development. Computational and Structural Biotechnology Journal. 2019;18:241. doi:10.1016/j.csbj.2019.12.006
2. GeeksforGeeks. Learning Rate in Neural Network. GeeksforGeeks. Published July 31, 2023. Accessed July 14, 2026. https://www.geeksforgeeks.org/machine-learning/impact-of-learning-rate-on-a-model/
3. Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010;50(5):742-754. doi:10.1021/ci100050t
4. Singh S, Kumar R, Payra S, et al. Artificial Intelligence and Machine Learning in Pharmacological Research: Bridging the Gap Between Data and Drug Discovery. Cureus. 2023;15(8). doi:10.7759/cureus.44359
5. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15:1929-1958. Accessed July 14, 2026. https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf
6. Olivia Harris, Michael Andrews. Effects and results of dropout layer in reducing overfitting with convolutional Neural Networks (CNN). World Journal of Advanced Engineering Technology and Sciences. 2024;13(2):836-853. doi:10.30574/wjaets.2024.13.2.0584
7. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. Journal of The Royal Society Interface. 2013;10(79):20120835. doi:10.1098/rsif.2012.0835
8. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the Decline in Pharmaceutical R&D Efficiency. Nature Reviews Drug Discovery. 2012;11(3):191-200. doi:10.1038/nrd3681

From “Being Seen” to “Being Recognized”: Content Presentation and Audience Resonance among Visually Impaired Creators on Short-Video Platforms By Zitian Yang

Abstract

Enabled by the platform economy and accessibility technologies, short-video creation has become a potential form of employment for visually impaired people. This study examines the most-liked videos of ten prominent visually impaired creators on Douyin, 200 highly liked comments, and the full comment corpus using case analysis and TF-IDF. The findings show that audience responses extend beyond sympathy to recognition of creators’ professional competence, independent living, and social participation, while also highlighting accessible facilities and public services. Short-video platforms may therefore enhance both visibility and capability-based recognition and support more inclusive employment.

I. Background

Visually impaired people continue to face substantial barriers in the labor market. Rather than arising solely from individual impairments, these disadvantages are often shaped by inaccessible workplaces, mobility constraints, limited occupational options, and inadequate organizational support (Schur et al., 2009; World Health Organization, 2022). Traditional jobs may impose strict requirements regarding vision and on-site work, thereby narrowing career choices and limiting opportunities for advancement. In China, the 2024 statistical bulletin reported 9.144 million employed persons with certified disabilities, including 2.787 million in flexible employment, underscoring both the scale and policy importance of disability employment support (China Disabled Persons’ Federation, 2025). Nevertheless, employment quality, income stability, and career development remain important concerns.

With the rapid growth of the internet economy and short-video platforms, together with the wider availability of accessible digital devices and assistive software, new forms of flexible employment have emerged. Short-video content creation has become one such occupation, reducing some constraints associated with physical mobility and fixed workplaces. Digital media can expand opportunities for self-expression and participation, although accessibility remains shaped by platform design and institutional support (Ellis & Kent, 2011). In this study, the ten selected visually impaired creators on Douyin had an average of 1.042 million followers; the most-followed creator, Dong Lina, had 4.533 million. These figures indicate that this form of platform-based work has attracted considerable public attention.

Compared with many conventional occupations, content creation offers distinctive opportunities for visually impaired people to present their experiences and skills directly to the public. Because sighted audiences may have limited exposure to the everyday lives of visually impaired people, creator-produced videos can generate curiosity, emotional engagement, and public discussion. The most-liked videos in the sample received an average of 1.465 million likes and 62,000 comments, indicating high levels of audience interaction. Independent digital content creation has therefore become a practical career option for some visually impaired people and

may help expand their employment opportunities. Examining the content and audience responses associated with this form of work is consequently of clear practical significance.

For a long time, visually impaired people have often been perceived primarily as recipients of assistance and social care, while their lived experiences, work capabilities, and career development needs have received limited public attention. Digital platforms and online storytelling can enable marginalized groups to present counter-narratives and participate more actively in public communication (Trevisan, 2017). At the same time, platform-based cultural production is shaped by platform infrastructures, visibility mechanisms, and commercial incentives (Abidin, 2016; Nieborg & Poell, 2018). Through short videos, visually impaired individuals can document their lives, share their experiences, showcase their talents, and interact with the public, thereby moving from being merely objects of public attention to becoming active agents of self-expression and content production.

In the title of this study, "being seen" refers to the process through which visually impaired creators use short-video platforms to present their living conditions, travel experiences, work activities, talents, and social interactions to a broader public. A typical example is a video by "Nuozi Miaowu," one of the sampled creators, documenting her independent air travel; the video had received 2.929 million likes and 75,000 comments by the time of data collection. It enabled viewers to observe how a visually impaired person navigates long-distance travel independently. Sighted viewers may otherwise have limited understanding of how visually impaired people travel independently, use guide dogs or assistive devices, pursue education and work, or encounter difficulties caused by occupied tactile paving and inadequate public services. By presenting these everyday experiences in a concrete, intuitive, and interactive manner, short videos bring them into public view and enhance the visibility of visually impaired people in public communication.

However, attracting public attention does not necessarily amount to gaining public recognition. If responses to visually impaired creators remain largely confined to emotional reactions such as feeling "moved," "distressed," or "inspired," or if creators are regarded mainly as "people in need of help," visually impaired individuals may still be positioned as objects of sympathy and care. In this study, "being recognized" refers to a broader form of public understanding: while acknowledging the practical difficulties faced by visually impaired people, audiences also recognize their capacities for independent living, content creation, professional expression, social interaction, and employment, and regard them as equal social actors with occupational value and career development needs. The most representative case is Dong Lina, the most-followed visually impaired creator in the sample, with 4.533 million followers. Her graduation speech video had received 3.454 million likes, and highly liked comments such as "This voice is definitely of national-broadcast quality" and "Her professional ability in broadcasting is outstanding" reflect recognition of her professional development and occupational value. For visually impaired people engaged in emerging forms of employment, extending audience responses from emotional attention to capability-based recognition is particularly important. Only when the public and platforms acknowledge not only their

difficulties but also the content and value they create is platform visibility more likely to translate into sustained employment support.

Against this background, this study examines the most-liked videos posted by ten visually impaired creators on Douyin and the comments associated with these videos. Through a case analysis of highly liked comments and a TF-IDF-based keyword analysis, it investigates the major life- and work-related scenarios presented by visually impaired creators, the types of content that generate audience resonance, and whether such resonance extends beyond emotional support to capability recognition, accessible environments, and employment support. In doing so, the study seeks to provide practical insights into how short-video platforms may support emerging forms of employment for visually impaired people and facilitate a transition from "being seen" to "being recognized."

II. Research Design and Methods

1. Research Platform

This study takes Douyin as its research setting. Douyin is a digital platform centered on short-video content production, dissemination, and user interaction. Users can engage in content creation and social interaction through video posting, liking, commenting, livestreaming, and other platform functions. The platform not only provides a space for users to present their everyday lives and express their views, but also creates opportunities for employment and entrepreneurship through content production, livestream interaction, product promotion, and personal brand building. These opportunities are nevertheless shaped by platform infrastructures and visibility mechanisms (Nieborg & Poell, 2018).

2. Sample Selection

This study selects ten visually impaired creators on Douyin as its research subjects and uses the most-liked video on each creator's account as the analytical sample. All ten were among the most-followed visually impaired creators on the platform. Account-level indicators, video metrics, and comments were collected in August 2025. The creators had an average of 1.042 million followers, with the highest follower count reaching 4.533 million (Dong Lina) and the lowest being 234,000 (Xiaguo Xiaguo). They had accumulated an average of 11.195 million total likes, indicating substantial influence in visually impaired content creation. The most-liked videos were selected because they generally achieved high levels of platform exposure and user interaction and therefore provide a concentrated view of the content most likely to attract public attention, recognition, and discussion.

This study adopts a purposive sampling strategy oriented toward platform reach. Its objective is neither to describe the average characteristics of all videos posted by visually impaired creators nor to estimate their overall income or employment performance on the platform. Rather, it seeks to identify the content scenarios most likely to generate audience resonance among videos with relatively broad reach and to examine what such resonance implies

for understanding platform-based employment among visually impaired people. Information on the selected creators and videos is presented in Table 1.

3. Research Methods

(1) Case Analysis

This study first employs a qualitative case-analysis approach to examine the twenty most-liked comments under each of the ten sampled videos, resulting in a total of 200 highly liked comments. These comments are not statistically representative of all audience responses, but they capture expressions that received relatively broad endorsement and visibility within the comment sections. The most-liked comment in the sample received 980,402 likes, while the top twenty comments under each video received an average of 58,230 likes.

Compared with randomly selected comments, highly liked comments are more likely to capture forms of expression that receive broader audience endorsement and stronger emotional engagement within the comment sections. They are therefore useful for identifying prominent themes of audience resonance.

More specifically, this study takes each individual comment as the basic unit of analysis and classifies and codes highly interactive comments by considering both the content of the corresponding video and the semantic meaning of the comment. Through this categorization, the study identifies whether audience resonance arises from sympathy for difficult circumstances, appreciation of social kindness, concern about accessibility governance, or recognition of visually impaired creators' capabilities and identities as professionals and workers.

(2) TF-IDF Keyword Analysis

Building on the case analysis, this study further applies the TF-IDF method to extract keywords from all comments collected under the ten sampled videos and to identify terms with relatively high thematic importance and distinctiveness in the corpus. TF-IDF, or term frequency–inverse document frequency, is a widely used term-weighting method in text analysis. Its basic logic is that a term is more informative when it appears frequently in a particular document but less frequently across other documents. By considering both within-document frequency and prevalence across the corpus, TF-IDF reduces the likelihood that common but thematically uninformative words will be identified as core keywords (Salton & Buckley, 1988).

The standard formula is expressed as follows:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log[N/\text{DF}(t)]$$

where t denotes a given term, d denotes a text unit, TF(t,d) represents the frequency of term t in text unit d, N represents the total number of text units, and DF(t) represents the number of text units containing term t.

III. Case Analysis

To identify the types of short-video content produced by visually impaired creators that are most likely to generate audience interaction and emotional responses, this study first analyzes the themes of the most-liked videos posted by the ten selected creators and the most-liked comments under those videos. For each video, the twenty comments receiving the highest numbers of likes were selected, yielding a total sample of 200 highly liked comments. Although these comments do not represent the attitudes of all viewers, they capture expressions with relatively strong communicative reach and endorsement within the comment sections.

1. Major Scenarios in Most-Liked Videos and Themes of Audience Resonance

The major scenarios in the most-liked videos and the corresponding themes of audience resonance are presented in Table 2.

The sampled videos that achieved relatively broad dissemination mainly involve career development and capability demonstration, independent travel and social participation, accessible environments and rights protection, benevolent interactions and social support, and assistive technologies and platform-based interaction. These scenarios not only present the everyday lives of visually impaired people but are also closely related to their participation in employment, content creation, and social life.

As shown in Table 2, the content that generates audience resonance is not limited to the difficulties faced by visually impaired people. Videos demonstrating professional capabilities, independent travel, and the use of technology also receive substantial attention. This suggests that audience responses extend beyond a sole focus on hardship to include recognition of independent living, professional competence, and social participation. At the same time, accessible facilities, public transportation services, and social assistance are frequently discussed in relation to the most-liked videos. The participation of visually impaired people in emerging forms of employment therefore depends not only on individual willingness and platform opportunities but also on mobility, the urban environment, public services, and technological support.

2. Themes of Public Resonance, Typical Expressions, and Employment Implications in Highly Liked Comments

Across different video scenarios, highly liked comments reveal a relatively clear set of themes of audience resonance, as presented in Table 3. Overall, the comments primarily concern professional competence, independent living and labor participation, accessible environments, benevolent interactions, assistive technologies, and emotional support.

The highly liked comments reveal a layered pattern of audience response. First, viewers express emotional reactions to the mobility difficulties and everyday inconveniences faced by visually impaired creators. Second, some comments address practical support conditions, such as tactile paving, public transportation, staff assistance, and assistive technologies. Third, in videos featuring career development, singing performances, and content creation, viewers more directly recognize the professional competence and occupational value of visually impaired creators.

These findings suggest that short-video platforms can do more than enhance the social visibility of visually impaired people; they can also provide channels through which creators demonstrate their capabilities, build public recognition, and expand employment possibilities. In particular, comments referring to "a master's graduate in broadcasting," "meeting clients," and "professional capabilities" indicate that visually impaired creators are not viewed solely as recipients of assistance or objects of sympathy. They are also recognized as professionals and workers who can learn, create, and provide value.

Nevertheless, strongly emotional expressions such as "moved," "heartbroken," and "inspirational" remain prominent in the highly liked comments. Such responses can attract attention and facilitate dissemination; however, if platform visibility relies heavily on difficult experiences and their emotional effects, visually impaired creators may remain confined to an image of people "in need of help." Therefore, while maintaining social care and concern, support for platform-based employment should place greater emphasis on content quality, professional competence, occupational value, and career development, thereby encouraging capability-based recognition and substantive employment support.

IV. Keyword Analysis

Building on the case analysis of highly liked comments, this study applies the TF-IDF method to extract keywords from the full corpus of comments collected under the sampled videos. The results are presented in Table 4. Whereas the case analysis focuses on typical expressions that generate strong user interaction, the keyword analysis identifies prominent objects of attention, situational contexts, emotional attitudes, and public issues within the comment texts. TF-IDF identifies terms with relatively strong thematic distinctiveness in the corpus; it does not directly indicate the proportion of comments expressing a particular attitude. To improve interpretability, the extracted keywords are classified, according to their meanings and contexts, into six categories: subjects and roles, scenarios and facilities, emotions and attitudes, interactions and behaviors, social and value-related issues, and tone and rhetorical expression.

As shown in Table 4, the keywords in the comment sections are distributed mainly across these six categories. Terms such as "blind person," "creator," "staff member," and "friend" indicate that discussions largely center on visually impaired creators and their interactions with helpers and service personnel. Terms such as "tactile paving," "guide dog," "white cane," "electric bicycle," "metro," and "city" suggest that mobility in public spaces and accessible environments constitute important areas of audience attention.

In terms of emotional expression, keywords such as "moved," "keep going," "warmth," "kindness," "brave," "heartbroken," and "aggrieved" are relatively prominent, indicating strong emotional resonance in audience responses. On the one hand, viewers express appreciation for visually impaired people's independent living, courageous mobility, and active self-expression. On the other hand, when videos involve occupied tactile paving, impeded travel, or unfriendly behavior, comments convey distress, dissatisfaction, and expectations of basic respect. These

findings are broadly consistent with the emotional support and value judgments identified in the case analysis.

Notably, the keywords are not confined to individual emotions. Terms such as “society,” “service,” “staff member,” “occupation,” “wrong-way riding,” “problem,” “convenience,” and “work” indicate that viewers connect the experiences of visually impaired creators with urban accessibility, public-service quality, and social inclusion. For visually impaired people, unobstructed tactile paving, convenient public transportation, and appropriate staff assistance affect not only everyday life but also independent travel, content creation, and participation in emerging forms of employment.

Overall, the TF-IDF keyword analysis and the case analysis of highly liked comments reinforce each other. Audience resonance arises not only from emotional responses to the difficulties faced by visually impaired creators but also from recognition of their independence and occupational value, and it extends to accessible environments and social support. Short-video platforms may therefore enhance social visibility while providing a channel through which the public can better understand the employment needs of visually impaired people. Future efforts should continue to move audience attention beyond being merely “moved” or “heartbroken” toward sustained recognition of professional competence, occupational value, and equal employment rights.

V. Conclusions and Recommendations

This study examines the most-liked videos posted by ten visually impaired creators on Douyin and the comments associated with these videos. Through a case analysis of highly liked comments and a TF-IDF-based keyword analysis, it investigates the audience resonance generated by the content presented by visually impaired creators. The findings show that the most prominent scenarios involve professional competence and personal development, independent mobility and social participation, accessible environments and rights protection, benevolent interactions and social support, and the use of assistive digital technologies. Audience responses also exhibit a clear thematic structure: viewers express sympathy and encouragement in response to practical difficulties, recognize professional competence and independent mobility, and connect individual experiences with broader public issues such as obstructed tactile paving, public transportation, service support, and social inclusion.

These findings indicate that short-video platforms can provide visually impaired people with a level of social visibility that is difficult to achieve through conventional media representation. Visually impaired creators are not merely objects represented in news reports or charitable narratives; they are content producers who document their lives, express their views, demonstrate their capabilities, and interact directly with the public. Through these videos, viewers can develop a more concrete understanding of the mobility, employment, and social-interaction needs of visually impaired people while also recognizing their professional competence, expressive capacity, and occupational value.

Based on these findings, this study proposes the following recommendations.

First, platform accessibility functions and support mechanisms for visually impaired creators should be improved. Short-video platforms should further optimize accessibility features for voice-based operation, video uploading and editing, livestream management, comment reading, and data monitoring, thereby lowering technical barriers to content creation and account operation. Visually impaired users with creative aspirations and employment needs could also be provided with digital-skills training, algorithmic visibility and promotional support, operational guidance, and opportunities for commercial collaboration, helping them move from occasional exposure toward more stable content creation and employment.

Second, public services and rights protection for persons with disabilities engaged in emerging forms of employment should be strengthened. Government agencies and disability-service organizations may incorporate short-video production, livestreaming, e-commerce operations, and online services into employment training and support systems while providing assistive devices, technical training, employment counseling, and legal protection. Attention should also be paid to income stability, labor protection, online rights violations, and the excessive emotional consumption of disability-related content, so that platform visibility does not create new forms of precarity for visually impaired creators.

Overall, the transition from “being seen” to “being recognized” describes both the career-development process of visually impaired creators on short-video platforms and the broader improvement of platform mechanisms, public services, and social perceptions. Only when society not only watches their stories but also recognizes their capabilities, supports their work, and improves their employment conditions can short-video platforms become an important channel for dignified and sustainable platform-based employment.

Works Cited

Abidin, C. (2016). Visibility labour: Engaging with influencers' fashion brands and #OOTD advertorial campaigns on Instagram. Media International Australia, 161(1), 86–100. https://doi.org/10.1177/1329878X16665177

China Disabled Persons' Federation. (2025). Statistical bulletin on the development of the work for persons with disabilities in 2024 [in Chinese].

Ellis, K., & Kent, M. (2011). Disability and new media. Routledge.

Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. New Media & Society, 20(11), 4275–4292. https://doi.org/10.1177/1461444818769694

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Schur, L., Kruse, D., Blasi, J., & Blanck, P. (2009). Is disability disabling in all workplaces? Workplace disparities and corporate culture. Industrial Relations: A Journal of Economy and Society, 48(3), 381–410. https://doi.org/10.1111/j.1468-232X.2009.00565.x

Trevisan, F. (2017). Crowd-sourced advocacy: Promoting disability rights through online storytelling. Public Relations Inquiry, 6(2), 191–208. https://doi.org/10.1177/2046147X17697785

World Health Organization. (2022). Global report on health equity for persons with disabilities. World Health Organization.

Table 1. Information on the Selected Creators and Videos (M = million; K = thousand)

Source Link	Creator Name	Followers	Total Likes	Video Title	Video Release Time	Video Duration	Video Likes	Video Comments	Video Favorites	Video Reposts
ttps://www.douyin.com/video/7320918559972953384	诺子喵呜	1.932M	3.817M	"If one day you lost your sight, would you have the courage to travel by plane on your own?" #AirTravel #Accessibility #BlindPeople #AsianParaGames	2024/1/6 17:43	10:35	2.929M	75K	109K	58K
ttps://www.douyin.com/video/7359159999517478207	宝哥和他的导盲犬	1.017M	7.312M	How inclusive is a civilized city toward guide dogs and blind people? #AccessibleTravel #GuideDogs #VisualImpairment #WarmthOfGuangzhou	2024/4/18 18:59	6:42	1.554M	48K	53K	08K
ttps://www.douyin.com/video/7274162439132826932	浩然	313K	.569M	You feel as comforting as curling up under a blanket, yet as elusive as the wind. #A_CappellaWithoutSoundCard #CoverSong #RedHighHeels #LoveMusicLoveLife #MusicianCreationProject @DOU+Assistant	2023/9/2 17:45	1:01	1.156M	83K	89K	94K

ttps://www.douyin.com/video/7452346018307263756	谭走鱼尾纹	309K	.820M	The previous post had limited reach, so I am reposting it. Sorry about that. #EverydayLife #BlindPeople #StreetStallDiary #Kindness #GoodSamaritans	2024/12/25 21:49	7:15	1.259M	102K	58K	55K
ttps://www.douyin.com/video/7455244000568675634	盲人博士生-黄莺	873K	6.562M	First-person travel diary by a blind creator: buying Mengmeng a birthday cake to express my affection ❤️ #VisualImpairment #BlindTravel #Accessibility #TactilePaving	2025/1/2 17:14	3:17	730K	24K	24K	0K
ttps://www.douyin.com/video/7382055954419911987	夏果夏果	234K	.893M	A day in the life of a blind university student. Although something unexpected happened, I still got to enjoy some delicious food. #VisualImpairment #Accessibility #DailyVlog #PositiveEnergy #Healing	2024/6/19 11:47	13:27	581K	24K	38K	3K
ttps://www.douyin.com/video/7463405873466051855	丸子小仙呀	294K	.803M	Mom, I'm home! A blind young woman's 1,200-km Spring Festival journey home. Thank you to every staff member and kind stranger who helped me along the way! #BlindPeople #Heartwarming #SpringFestivalTravelRush #Homecoming #PositiveEnergy #KeepFilmingForTheNewYear	2025/1/24 17:07	14:42	2.556M	78K	91K	09K
ttps://www.douyin.com/video/7447080855333375273	蒙眼炒鸡蛋	610K	.245M	I almost got myself lost on the metro today /(ㄒoㄒ)/~~ Many thanks to the kind man and the staff who helped me. #BlindPeople #BlindTravel	2024/12/11 17:17	6:03	311K	12K	23K	6K
ttps://www.douyin.com/video/7453880729097751808	盲人橘子	306K	.186M	This little guy is just too adorable! #BlindPeople #PositiveEnergy #Trending	2024/12/30 6:00	3:58	123K	4,700	2,753	,250
ttps://www.douyin.com/video/7249590627464236345	董丽娜	4.533M	9.744M	I was so excited during my graduation speech. Today, I graduated with everyone! Thank you to the university for giving me the opportunity to speak on behalf of the graduating class. Best wishes to the Class of 2023 for a happy graduation and a bright future. #Broadcasting #GraduationCeremony #University #Speech #GraduationSeason	2023/6/28 12:34	8:42	3.454M	171K	322K	72K

Table 2. Major Scenarios in Most-Liked Videos and Themes of Audience Resonance

Video Scenario Category	Representative Video Situations	Main Focus of Highly Liked Comments	Relevance to Employment among Visually Impaired People
Professional Development and Capability Demonstration	Visually impaired people completing higher education and delivering graduation speeches; visually impaired singers performing	Professional competence, vocal performance, challenging stereotypes, and self-realization	Helps the public view visually impaired people as workers with professional competence and career potential
Independent Mobility and Social Participation	Traveling alone by plane, taking the metro with a guide dog, and traveling independently between cities	Courage, independence, digital tools, and public-service support	Mobility and service accessibility are essential prerequisites for employment and content creation

Accessible Environments and Rights Protection	Occupied tactile paving, white canes being struck, and accessible facilities on buses and roads	Governance of tactile paving, public facilities, basic respect, and awareness of rules	Accessible environments directly affect commuting, work, and social participation
Benevolent Interaction and Social Support	Assistance from staff, guidance from passers-by, and coordinated railway-service support	Warmth, assistance, civic-mindedness, and social inclusion	Social support can reduce the practical barriers to participation in employment and public life
Assistive Technology and Platform Interaction	Mobile-phone accessibility mode, suggestions for smart canes, and online assistance tools	Technological improvements, digital tools, and assistance through livestream connections	Digital technologies and platform functions may expand opportunities for employment and content creation

Table 3. Themes of Public Resonance, Typical Expressions, and Employment Implications in Highly Liked Comments

Theme of Public Resonance	**Illustrative Excerpts from Comments**	**Public Attitudes Reflected in the Comments**	**Implications for Emerging Forms of Employment among Visually Impaired People**
Professional Competence and Personal Growth	"From a blind massage therapist to a visually impaired master's graduate in broadcasting"; "Music will never reject you"	The public recognizes the professional competence and expressive value of people with disabilities	Platforms should provide greater support for skill demonstration, knowledge sharing, and professional content creation
Independent Mobility and Social Participation	"Traveling hundreds of kilometers entirely on her own by plane"; "Taking the metro to meet clients as part of daily life"	Viewers recognize visually impaired people's ability to act independently and participate in social life	Accessible transportation and digital services are essential prerequisites for platform-based employment
Accessible Facilities and Rights Protection	"What they need is the most basic respect"; "Bus-stop assistance buttons should be introduced nationwide"	The public reframes individual experiences as matters of public governance and rights	Accessible environments should be incorporated into employment-support systems
Benevolent Interaction and Public Services	"A heartwarming relay of assistance across 1,200 kilometers"; "This city is so warm"	Viewers value the friendly support provided by public-service staff and members of the public	Informal assistance should be transformed into routine, standardized service mechanisms
Assistive Technology and Platform Support	"Smart guide cane"; "The phone's accessibility mode is very useful"; "Providing guidance through livestream connections"	The public expects technology to reduce barriers in daily life and work	Platforms and technology companies should strengthen the development of accessible tools for content creation and employment
Emotional Resonance and Dissemination	"It made me cry"; "This is heartbreaking"; "How many more times does she have to say 'thank you' than the rest of us?"	The public expresses strong sympathy and emotional support	Employment support should avoid overreliance on narratives of hardship and promote capability-based recognition

Table 4. Classification of TF-IDF Keywords in the Comment Sections of the Sampled Videos and Their Core Meanings

Category	**Keyword Examples**	**Core Meaning/Explanation**
Subjects and Roles	blind person, staff member, young woman, creator, friend, kind person, person with a disability, police officer	Comments revolve around the people appearing in the videos, particularly blind people and related helpers, bystanders, and media figures
Scenarios and Facilities	tactile paving, sidewalk, guide dog, Alpha, white cane, electric bicycle, metro, city, Xi'an, Guangzhou, food delivery, work, place	Reflects mobility, urban accessibility, and everyday settings—the spatial and institutional contexts that attract viewers' attention
Emotions and Attitudes	moved, keep going, thank you, warmth, kindness, brave, cute, beautiful, like, pleasant-sounding, impressive, gentle, hope, heartbroken, aggrieved, tears, don't, can't	Highly emotional language dominated by sympathy, encouragement, and expressions of being moved, while also conveying distress and opposition (e.g., "don't" and "can't")
Interactions and Behaviors	see, cannot see, unable to see, catch sight of, help, encounter, apologize, like, comment, edit, know, think, watch	Captures interactions between blind people, the public, and commenters; includes perceptual and expressive actions typical of social-video discourse
Social and Value-Related Issues	society, country, China, service, staff member, obstruction, wrong-way riding, problem, convenience, need, hope, life, work, service, well-being	Reflects public discussion of social inclusion, institutional improvement, and civil conduct in public spaces, constituting the public-issue dimension of the topic
Tone and Rhetorical Expression	really, properly, somewhat, especially, think, seems, actually, certainly, later, now, very, this kind of, a little, every, suddenly, already, like	Typical colloquial comment language that intensifies emotional tone and forms the "empathic register" commonly found on short-video platforms

Predicting Bearing Failure 48 Hours in Advance: A Hybrid CNN-LSTM-Attention Approach on the IMS Dataset By Matt Wang

Research question: Can a machine learning model, integrating multi-sensor time-series data, achieve ≥90% precision in predicting the remaining useful life or imminent failure of an industrial robot at least 48 hours in advance?

Biography

Yifu Wang (Matt) is a High School Student from English Schools Foundation South Island School, part of the graduating class of 2027. He competes in High School Division Robotics, and indulges in playing Go and the Piano in his free time. Matt is deeply fascinated by the intersection of data science and robotics, and how it can be applied to real world problems. Matt plans to pursue a degree in Electrical or Industrial Engineering and minor in Computer Science, hoping to bridge predictive AI with industrial automation.

Abstract

Unplanned machine failures can cause interruptions in production and raise maintenance costs, but automated monitoring is vulnerable to false alarms. This study poses the question of whether a machine learning model, integrating multi-sensor time-series data, can achieve ≥90% precision in predicting the remaining useful life or imminent failure of an industrial robot at least 48 hours in advance. In answering this question, vibration recordings from four run to failure experiments in the Intelligent Maintenance Systems (IMS) bearing dataset, were extracted and converted into three indicators of degradation, namely root mean square amplitude, peak amplitude, and crest factor. Each of these indicators were normalized against the healthy baseline in their respective runs. A hybrid deep learning model which combines a convolutional branch, a recurrent branch, and a temporal attention mechanism, was trained to output the failure probability for the next 48 hours as well as an estimate of the remaining useful life (RUL). Two bearing runs were used for training, one independent run was used for model selection and tuning the decision threshold, and one fully held out run was used for evaluation. During final evaluation on the held out run, the model raised 0 false alarms across 5,780 healthy windows, all while recalling 52.6% of true prefailure windows, predicting RUL with a mean absolute error (MAE) of 23.3 hours. However, the first sustained alert arrived only 25.2 hours before failure, short of the 48 hour target, but still enough time for intervention to be scheduled. These results show that very high precision failure warning is attainable from just a small set of vibration data, but the achievable lead time is limited by how individual bearings show signs of degradation, rather than the model itself.

Keywords

Machine learning, predictive maintenance, IMS, sensor degradation, fault diagnosis,

Introduction

Nearly all industrial production requires the support of rotational machinery. Among its most failure prone components, is the rolling element bearings. Bearings support the shafts of motors, gearboxes, and joint drives of industrial robots. The degradation of bearings is a leading cause of unplanned downtime (Jardine et al.). Even one unassuming bearing failure can halt an entire production chain, and thus maintenance practice has moved beyond simply repairing equipment after it breaks, towards trying to predict failures before they even happen (Mobley). The value of such predictions depend on the accuracy of two values, which are the warning lead time, determining whether replacement parts and repair can be arranged in advance, and the precision of the alarms, which determine whether operators trust the system enough to act on alarms at all. The research question of this study asks whether a machine learning model, integrating multi sensor time series data, can achieve at least 90 percent precision in predicting the remaining useful life or imminent failure of an industrial machine at least 48 hours in advance. My hypothesis is that a hybrid deep learning model combining convolutional, recurrent, and attention components, trained on vibrational features normalized to each machine's own healthy baseline, can satisfy requirements on a run to failure it had never seen before. To explore the research question, the hypothesis was tested using the Intelligent Maintenance Systems (IMS) bearing dataset under strict evaluation protocols, where the test bearing comes from a different experimental campaign, rather than every bearing being used for training and tuning. The decision threshold was frozen before the test data was analyzed, and the lead time was reported as a primary outcome alongside precision.

Literature Review

Maintenance strategy has developed through three distinct stages. The earliest approach to repair was simply run to failure maintenance, repairing equipment and machinery only after it breaks. This maximizes the use of each individual component, but at the cost of unplanned downtime and secondary damage to machines. The second stage was preventive maintenance, where machine servicing was scheduled at fixed time intervals. This avoids many failures, but wastes component life and labor on machines that were still healthy (Mobley). Moreover, parts that weren't subject to regular replacement may still break down unexpectedly, causing unplanned downtime. The third stage is condition based and predictive maintenance. This is where sensor measurements are monitored continuously, scheduling interventions only when data analysis indicates that degradation has begun (Jardine et al.). Previous empirical work has linked proactive maintenance strategies to measurable gains in equipment performance and cost, leading to sustained industrial interest in making failure prediction reliable enough to implement.

Within the field of predictive maintenance, rolling element bearings, in particular, have generated major attention, because they are ubiquitous, inexpensive individually, and are disproportionately responsible for machine breakdowns. Bearings degrade through recognizable physical stages. A microscopic subsurface crack can grow into a large surface defect on a race or roller, generating periodic impacts as rolling elements strike it, spreading damage until the

bearing can no longer support its load (El-Thalji & Jantunen). These impacts leave signatures in the vibration signal measuring in the bearing, which is why vibration analysis has become the dominant sensing approach for bearing condition monitoring (Randall & Antoni).

Classical bearing diagnostics reduces the vibration signal to hand crafted statistical indicators and tracks them over time. Root mean square amplitude reflects the overall vibration energy, while peak amplitude measures the size of individual impacts. Kurtosis and crest factor measure how impulsive the signal is relative to its background level, and envelope analysis can isolate the characteristic frequencies of specific defect types. These indicators are physically interpretable and cheap to compute, but using them traditionally requires experts to set alarm thresholds by hand, and a fixed threshold on a single indicator would ignore how the indicators change together over time. Work on weak signature detection has shown that early stage defects can be buried in noise and remain invisible until damage is well developed, limiting how early a threshold based system can send a warning (Qiu et al.).

To alleviate these limitations, data driven prognostics started developing, where models learn the relationship between sensor measurements and future failure using run to failure sensor data. Reviews of this field distinguish between diagnostic tasks, which identify a fault that already exists, and prognostic tasks, which estimates when a machine will fail. This is commonly represented as the remaining useful life, or RUL (Sikorska et al.). Early on, machine learning approaches applied classifiers like support vector machines and random forests, to extracted features, which demonstrated that learned models could outperform fixed thresholds, all while balancing the competing costs of false alarms and missed failures (Susto et al.). Systematic reviews of predictive maintenance through machine learning have witnessed a steep increase in attention, but also a wide variation in evaluation practices, though most studies report that accuracy alone on data splits that don't actually reflect deployment conditions (Carvalho et al.).

Deep learning tended this line of work by removing manual features engineering and by modeling temporal structure directly. Convolutional neural networks apply learned filters across time axes and can detect local patterns such as sudden amplitude jumps wherever they occur in the signal. They were first applied to RUL estimation on benchmark datasets by Babu, Zhao and Li (Sateesh Babu et al.) and later convolutional architectures achieved substantially lower error rates on the same benchmarks (Li et al.). However, convolutional models alone have limited view of the long term data context, since each filter only sees one short segment of the signal at a time.

Recurrent networks address this limitation by carrying information forward through time. The long short term memory (LSTM) architecture adds information filters, allowing networks to retain information across very long sequences without the training instabilities found in earlier recurrent designs (Hochreiter & Schmidhuber). Recurrent networks were applied to RUL estimation as early as the first prognostics benchmarking competitions (Heimes). LSTM models were later shown to outperform both convolutional and classical baselines on standard prognostics datasets, because degradation is inherently a cumulative process, where the current health state depends on the entire history of operation (Zheng et al.).

Latterly, two further developments shaped the architecture used in present studies. The first is the attention mechanism, which allows a model to assign different weighings to specific time steps, rather than treating the whole monitoring window equally (Vaswani et al.). Attention based models for bearing prognostics can focus on segments of windows where degradation actually accelerates, and have been shown to have improved RUL prediction on bearing datasets (Chen et al.). The second development is the treatment of the regression target itself. Since vibration signals carry no actual information about RUL while a machine is still healthy, most benchmark practices cap the RUL target at a fixed value, producing a piecewise linear target that the model can realistically learn from (Saxena et al.).

However, despite the progress, a gap still remains between how these models are evaluated and what a deployed early warning system actually needs to do. Comprehensive reviews note that most studies report average regression error or classification accuracy. On the contrary, operational requirements are stated as a minimum precision and a minimum lead time, and that evaluation protocol frequently allows information from the test machine to influence training or threshold selection (Lei et al.). Average error can be low while the alarm still arrives too late to act on, and accuracy can be high whilst high false alarms make operators ignore the system.

The Intelligent Maintenance Systems bearing dataset, which records complete run to failure histories of physically separate bearings across three independent campaigns make a stricter protocol possible. Since entire campaigns can be held out from training. This study will use that dataset to evaluate a hybrid deep learning model against the precision and lead time requirements directly, with the decision threshold tuned on an independent run, and then frozen before examining the held out test run.

Methods

Dataset Information

The data is provided by the IMS at the University of Cincinnati, generated on a test rig with four bearings mounted on a shared shaft. Each record is a short high frequency vibration snapshot taken at regular intervals over a full run to failure. Although the rig is not an industrial robot, the rolling bearings simulate the kind of movement common in robot joints and drive systems, thus the dataset provides a controlled way to test whether multivariate vibration time series data can support early failure warning before a complete system dataset is available.

Vibration data were collected using PCB 353B33 High Sensitivity Quartz ICP accelerometers mounted on the bearing housings. Data acquisition was performed using an NI DAQ Card 6062E. Each data file represents a one second vibration signal snapshot. Each file contains 20,480 data points sampled at 20 kHz, which corresponds to approximately 1.02 seconds of waveform data. The file name records the time at which the data were collected. The dataset contains three separate experiments organised as separate folders. Because each

experiment ran on physically distinct bearing instances at different times, splitting by run guarantees that no vibration data from the test bearing was ever seen during training or tuning.

Run	Documented failure mode	Snapshot files	Duration (hrs)	Role
Set 1, Bearing 3	Inner race defect	2156	827.6	Training
Set 1, Bearing 4	Roller element defect	2156	827.6	Training
Set 2, Bearing 1	Outer race failure	984	163.8	Validation (Including model selection, threshold tuning)
Set 3, Bearing 3	Outer race failure	6324	1073.3	Test

Table 1: Run-to-failure experiments used in this study.

Even though the rig setup is a bearing test bench rather than an industrial robot, the rolling element bearings of this type are common in robot joints and drive chassis systems, so this dataset provides a controlled and openly reproducible setting to answer the research question.

Feature extraction and baseline normalization

Each 20,480 snapshot was reduced to three scalar condition indicators from a single accelerometer channel, including the root mean square (RMS) of the signal, the peak absolute amplitude, and the crest factor (peak amplitude divided by RMS). These indicators were chosen because they are standard and physically interpretable measures of bearing degradation, RMS reflects overall vibration energy, while peak amplitude and crest factor correspond to the impacts produced by localized surface defects (Randall & Antoni), and because using identical features for every run gives all inputs the same physical meaning across individual experiments.

Absolute vibration levels differ between bearings, sensors, and mounting conditions, hence each feature was divided by its median value over the first 24 hours of its own run, the period during which every bearing is healthy. After this normalisation, a value of 1.0 means “at the healthy baseline”, and for example, 3.0 would mean “three times the healthy level”. This representation transfers across runs. Timestamps were converted to elapsed hours from the start of each run.

Two label types were assigned to every snapshot. The binary failure label was 1 if the snapshot fell within 48 hours of the end of its run, otherwise 0. The failure time is taken as the final recorded file, since each campaign was stopped at failure. A secondary 72 hour label was created for a sensitivity experiment. The RUL label was the number of hours remaining until the end of the run, and a capped variant limited RUL to 130 hours. This follows the standard piecewise linear target used in prognostics benchmarks, which reflects that degradation is not observable in vibration signals while a bearing is still healthy (Saxena et al.).

Preprocessing and window construction

Rows were sorted chronologically within each run, standardized per feature using statistics fitted on the training runs only, and cut into overlapping sliding windows of 256 consecutive snapshots with a stride of one snapshot. Each window was labelled by the failure label and capped RUL of its final snapshot, so that each prediction only uses past and present information. This produced 3,802 training windows (330 positive, 8.7%), 729 validation windows (289 positive), and 6,069 test windows (289 positive, 4.8%). RUL regression targets were divided by the maximum training value (130h) for numerical stability and rescaled to hours for reporting. The full experimental workflow is summarised below in Figure 1.

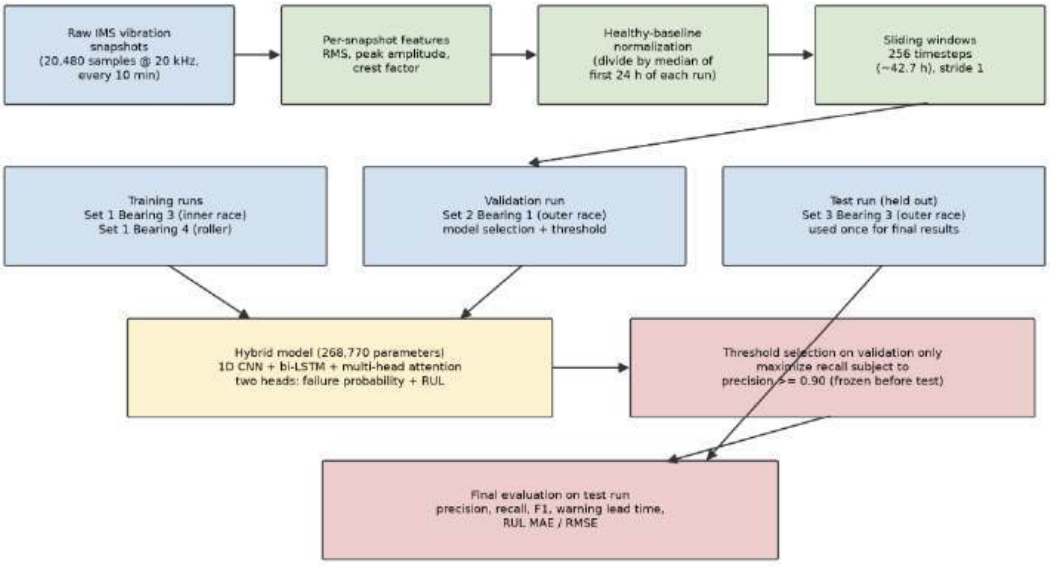


Figure 1: End-to-end experimental pipeline

Model Architecture

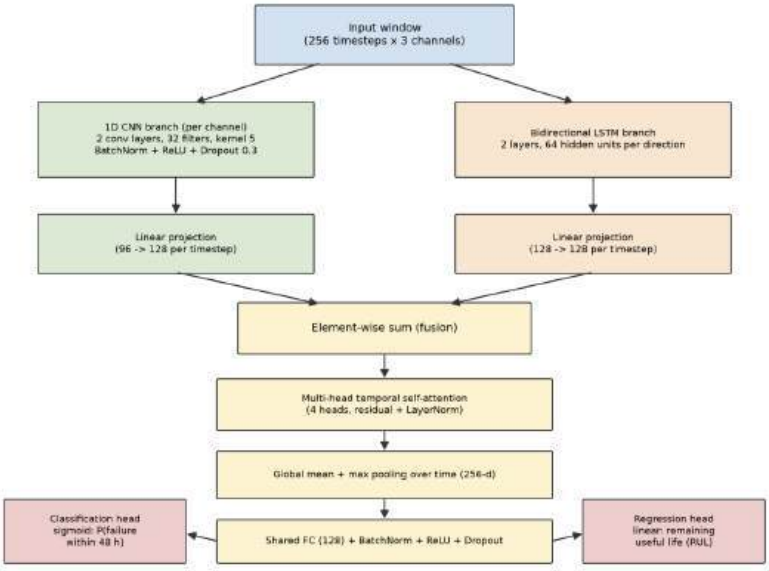


Figure 2: Hybrid CNN-LSTM + Attention architecture with two task heads

The model shown in figure 2, is a hybrid deep neural network with two parallel encoders, whose outputs are fused and refined by attention. A one-dimensional convolutional branch, with two layers, 32 filters of kernel size 5, each followed by the batch normalization, rectified linear activation, and dropout of 0.3, processes each input channel separately to extract local temporal patterns, such as sudden amplitude shifts.

A bidirectional LSTM branch, with two layers, 64 hidden units per direction, processes the full multivariate sequence to capture the slow monotonic drift of degradation. Both branches are linearly projected to a common 128-dimensional representation per time step and fused by element-wise addition. A multi-head self attention layer then reweights time steps, so the most informative parts of the 42.7 hour window occupies most of the final representation. The global mean and max pooling over time, followed by a shared fully connected layer of 128 units with batch normalisation, rectified linear activation, and dropout, feed two task specific heads, a sigmoid classification head which outputs the probability of failure in the next 48 hours, and a linear regression head that outputs RUL. The complete model has 268,770 trainable parameters.

The two tasks were trained jointly since they are complementary views of the same degradation process. The classification head returns a calibrated alarm signal, while the regression head provides an interpretable estimate of time to failure.

Training Procedure

The model was trained through minimizing combined loss that added the binary cross entropy from the classification head and the mean squared error from the regression head together, with both tasks weighted equally. One real concern here was class imbalance, as only

8.7 percent of the 3,802 training windows carried a positive label. If the loss treated both classes equally, the network may potentially reach a lower loss simply by predicting that no failure would ever occur. To prevent this, the positive term in the cross entropy was multiplied by the ratio of negative to positive training windows, coming out to approximately 10.5. With this weighting, one missed failure costs the model about as much as ten and a half false alarms, which pushes the optimizer toward learning the rare failure patterns rather than ignoring them. The remaining useful life targets were also rescaled before training. Each target was capped at 130 hours and then divided by that limit so the regression head learned values between zero and one, which kept the two loss terms on a similar numerical scale and stopped the regression error from dominating the gradient early in training. Reported errors were converted back into hours afterward so the results remain interpretable.

Optimization used the Adam (Adaptive Moment Estimation) algorithm (Kingma & Ba) with an initial learning rate of 0.001 and a batch size of 32 windows, for a maximum of 50 epochs. A plateau scheduler monitored the validation loss and halved the learning rate whenever five consecutive epochs passed without improvement, which allowed the model to take large steps early and finer steps once progress slowed. Dropout with a rate of 0.3 was applied inside the convolutional, recurrent, and fully connected blocks as regularization, which mattered because a network of roughly 269,000 parameters can easily memorize a training set built from only two run-to-failure trajectories. After every epoch the model was evaluated on the 729 windows of the held-out validation run, and only the checkpoint with the lowest validation loss was kept for later use. This early selection step acted as an additional guard against overfitting, since checkpoints from later epochs fit the training runs more closely but generalized worse to the unseen bearing. The random seed was fixed at 42 for Python, NumPy, and PyTorch so that the entire run can be reproduced exactly. Training was carried out in Python 3.12.0 with PyTorch 2.10.0, NumPy 2.4.2, pandas 3.0.1, scikit-learn 1.8.0, and SciPy 1.17.1 (Paszke et al.) on an Apple laptop CPU and took approximately 38 minutes to complete using the training.

Decision Threshold selection & Evaluation protocol

The research question asks whether the model can reach a precision of at least 90 percent, which shaped how the classification output was handled. The network outputs a continuous probability between zero and one for every window instead of a direct alarm. Some cutoff must be chosen above which the system declares that a failure is coming. That cutoff is a separate decision rule that's applied after training, separate from the model itself. This controls the exact tradeoff the research question poses. A high cutoff makes the system cautious, so it raises fewer alarms but the alarms it does raise are more likely to be correct, which favors precision. A low cutoff makes the system sensitive, so it catches more of the true failure windows but risks more false alarms, which favors recall. Accordingly, candidate thresholds from 0.05 to 0.95 in steps of 0.05 were each evaluated on the validation predictions, and the one selected was the threshold that produced the highest recall among all candidates whose validation precision was at least 0.90.

How the threshold was chosen matters, but where it was chosen matters too. If the threshold had been tuned on the test run itself, the reported precision would be circular, since the cutoff would have been adjusted after seeing the very answers it was being graded on, and the result could not honestly claim to predict anything. To avoid this, the threshold was selected using only the validation bearing, which is a physically separate run from a different experiment than the test bearing. Once selected, the threshold was frozen, and the held-out test run was evaluated exactly once at that fixed value.

Model performance was measured in three ways. The first was the quality of the 48 hour failure classification. Precision, recall, F1 score, and accuracy were computed using the confusion matrix of the test run, and precision was treated as the headline number because the research question sets a minimum precision of 0.90.

The second was the warning lead time, which measures how far in advance the model raised its first trustworthy alarm. An alarm was only counted once the predicted failure probability stayed above the threshold for three windows in a row. This rule exists because a single noisy window could briefly cross the threshold by chance, and a real maintenance team would not act on one isolated spike. The lead time was then defined as the true remaining useful life at the first window of that sustained alarm. The research question requires this value to be at least 48 hours.

The third was the accuracy of the remaining useful life estimate itself, measured with mean absolute error and root mean squared error. These were reported against both the raw and the capped RUL targets, and also separately for the final 130 hours of the run, since that is the period when an RUL estimate would actually influence a maintenance decision.

Results

Before any modeling, the prepared features were examined to verify that they encode degradation and to characterize the four runs.

Figure 3 shows the baseline normalized trajectories. Each four of the bearings show the expected patterns, which features near 1.0 for most of the run, then a rise toward failure. However, the shape of the rise differs sharply between each run. The two training bearings (Set 1) degrade noisily over the final 100-200 hours. The validation bearing (Set 2 Bearing 1) shows a clear, gradual incline beginning roughly 50 hours before failure. The test bearing (Set 3 Bearing 3) remains near baseline for over 1,000 hours and then degrades abruptly in the final roughly 30 hours. This heterogeneity is a known property of the IMS dataset (Qiu et al.) and, as shown in the figure, directly determines the achievable lead time.

In figure 4, we see the Pearson correlation matrix of three input features and labels on the training data. Each of the three features correlate negatively with RUL and positively with the 48 hour failure label, which confirms that each carries degradation signal. Furthermore, peak amplitude and crest factor are strongly correlated with each other (0.95), meaning that the effective feature set is closer to two independent signals than three. The moderate correlations

also indicate not one single feature separates healthy from failing on its own, which motivates the model that combines them over time.

Experiment 1: Training and threshold selection the validation run

The first experiment tested whether or not the hybrid model could learn a high precision alarm from only two training trajectories. The model reached its best validation loss (1,296) on the first epoch. Validation loss rose steadily afterward, peaking at 6.74 at epoch 36, while training loss continued to fall, showing a clear signature of overfitting to the two available training runs. The checkpoint selection protocol handled this as designed, retaining the model from epoch 1, but the observation is reported here as an important characteristic of the experiment. With only two run to failure trajectories from a single campaign, the 268,770 parameter model extracts all generalizable structure almost immediately, and further training only memorises run-specific noise. Since only two training trajectories were available, the results reported only reflect what is achievable with minimal training. Data availability is the limiting factor here, not the architecture.

Threshold	Precision	Recall	Selected
0.05	1.000	0.976	Yes (maximized recall, ≥0.90)
0.50	1.000	0.886	—
0.95	1.000	0.602	—

Table 2: Threshold selection on the validation run

On the validation run, the threshold search found that precision was 1.000 at every candidate threshold. Thus, the constraint precision ≥0.90 was satisfied every time. The selection rule chose the threshold maximizing recall, giving validation precision 1.000 and recall 0.976. However, the choice is insensitive. Figure 5 shows that the predicted probability stays near zero throughout the healthy period and rises steeply to near 1.0 as degradation begins, leaving few windows in the ambiguous middle region. The first sustained alert on the validation run occurred 46.8 hours before failure, almost just right at the 48 hour label boundary.

Experiment 2: Single evaluation on the held-out test run

The second experiment applied the finished system to the held-out test run. By this point, every adjustable part of the pipeline had been fixed in advance. The model weights were those of the checkpoint that achieved the lowest validation loss during training, and the alarm threshold of 0.05 was the value selected on the validation bearing in the previous experiment. Nothing about the test run was allowed to influence any setting before it was scored. The test run itself was held with Set 3 Bearing 3, consisting of 6,069 windows, and it comes from a different experimental campaign than every bearing used for training and validation, so it represents a genuinely unseen machine rather than a rearrangement of familiar data.

Metric	Value	Criterion	Outcome
Precision	1.000	≥ 0.90	Met
Recall	0.526	—	—
F1 score	0.699	—	—
Accuracy	0.977	—	—
True/false positives	152 / 0	—	—
False/true negatives	137 / 5780	—	—
Warning lead time	25.2h	≥ 48h	Not met
RUL MAE vs capped target	23.3h	—	——
RUL RMSE vs capped target	28.1h	—	
RUL MAE, final 130h	45.7h	—	—

Table 3: Final test run performance

The precision part of the hypothesis was strongly supported. At the frozen threshold, the model produced zero false positives across 5,780 healthy windows, giving a test precision of 1.000, recall of 0.526, F1 score of 0.689, and accuracy of 0.977(Table 3 & Figure 8). Figure 7 below shows that test precision was 1.000 at every threshold in the search grid, hence this result is not a result of the specific threshold chosen. The model's probability simply never rises above 0.05 while the test bearing is healthy, across more than 1,000 hours of monitoring.

However, the lead time half of the hypothesis was not supported on this bearing. The first sustained alert occurred just 25.2 hours before failure (Figure 6), and recall within the labeled 48 hours window was 52.6%. This shows that the model detected the second half of the pre-failure window, but had missed the first half. The zoomed panel of Figure 6 shows the cause of this directly. The predicted probability tracks the physical degradation signal, which on this bearing only departs from baseline roughly 30 hours before failure (compare Figure 3, bottom right). All 137 false negatives are windows in which the bearing was within 48 hours of failure, but had not yet begun to vibrate abnormally. Therefore, the model warned essentially as early as the chosen features made possible, though it remains a genuine negative result against the 48 hour criteria of the research question. This lead time result comes from a single held out trajectory, so it can't be generalised with statistical confidence. Under the same protocol, the validation run, which degraded more gradually, achieved 46.8 hours.

Figure 8 breaks the test result down into its four possible outcomes. Of the 5,780 windows where the bearing was genuinely healthy, none produced an alarm, and this empty false alarm cell is what drives the perfect precision score. The cost of that caution appears in the lower left cell, where 137 of the 289 failing windows went undetected. These missed windows are concentrated in the earlier part of the 48 hour failure zone, before the vibration signal had risen

far enough above baseline to be distinguishable from noise. Therefore, the matrix captures both halves of the finding in one picture. The model never raises a false alarm, but it cannot flag a failure before physical evidence of one exists.

Experiment 3: Remaining useful life regression

Because a binary alarm alone cannot tell an operator how urgently to act on a robot or machine, the third analysis assessed the regression head on the test run. Figure 9 compares predicted RUL with the capped ground truth. During the healthy period, the model predicts approximately 105-120 hours against the 130 hour cap, which is a stable underestimate of about 15-25 hours. During the final descent, the prediction falls in the correct direction, but lags the true countdown, overshooting below zero in the last hours after failure onset. The overall test MAE against the capped target was 45.7 hours. Compared against raw uncapped RUL, the MAE was 417.3 hours. This is expected as the model was trained on capped targets precisely because vibration features carry no information about RUL when the bearing is still healthy, hence it won't be able to distinguish between 1,000 from 130 remaining hours.

Sensitivity experiment 72 hour labeling horizon

Due to the lead time shortfall in Experiment 2, which might be an artifact of the 48 hour label definition, a variant model had to be trained identically but with positive labels covering the final 72 hours, and evaluated with the same frozen threshold protocol against the 72 hour labels. While the wider horizon didn't improve early warning, with test precision persisting at 1.000, recall fell to 0.403 (157 of 309 positive windows), and its first sustained alert moved marginally, to 26.0 hours before failure. This supports the interpretation from Experiment 2, that the binding constraint on lead time is the late onset of measurable degradation in the test's bearing vibrational signal.

Limitations and assumptions

There were several principal limitations to this study. Firstly, only four run to failure trajectories were available, two for training, one for validation, and one for testing. Hence all test- run conclusions rest on a single trajectory. Secondly, failure time was defined as the final recorded snapshot of each campaign, which seems conventional for this dataset, but only approximates the true moment of functional failure. Thirdly, the three features come from a single accelerometer channel per bearing, thus the "multi-sensor" integration is across derived indicators from one physical sensor, rather than across a heterogeneous variety of sensor types, like temperature or current. Fourthly, the model was selected at epoch 1, due to immediate overfitting, meaning that the training dataset limits performance. Lastly, the healthy baseline normalization assumes that the first 24 hours of a deployment are fault free. While this held for the runs that were assessed, they must be verified in any new application.

Discussion

All together, the experiment gives a clear and consistent answer to the research question: Can a machine learning model, integrating multi-sensor time-series data, achieve ≥90% precision in predicting the remaining useful life or imminent failure of an industrial robot at least 48 hours in advance? The precision requirement was met, on a held out run to failure experiment from an unseen dataset entry, the model raised 152 correct alarms and zero false alarms in more than 1,000 hours of monitoring, with its precision insensitive to the decision threshold across the entire search grid. However, the lead time requirement was only met on the validation bearing (46.8) hours, but not on the test bearing (25.2 hours). The sensitive experiment with a 72 hour label showed that the shortfall isn't due to label design. A recurring element was visible in the degradation trajectories themselves, where the model warned as soon as the vibration features left the healthy baseline, thus the achievable lead time was equal to the duration of each bearing's measurable degradation phase. Where degradation is gradual, the 48 hour target is within reach, but where a fault develops abruptly, no classifier operating on these features could have been alarmed earlier, without sacrificing the 1.000 precision that defines this system.

This tradeoff is the central finding of the experiment. It refines how the original hypothesis should have been understood. The hypothesis implicitly assumes that lead time is a property of the model. However the results show that it is jointly a property of the model, the features, and the failure physics of each individual robotic component. This is consistent with existing prognostic literature, which commonly distinguishes an incipient-fault detection horizon set by sensing physics, from the predictional quality achievable after that horizon (Lei et al.). Within the window that is physically allowed, the architecture performed close to the ceiling, though recalling labeled prefailure windows was limited almost entirely by windows that contained no degradation signal yet. The RUL head tracked the final countdown with a mean error of roughly one day, adequate for scheduling an intervention.

The practical implication for predictive maintenance has a caveat. While a system with near perfect precision and roughly one day of warning is already operationally valuable, the caveat is that a hard guarantee of 48 hours or even 24 hours cannot be promised for every failure mode from the vibration summary statistics alone. Extending the range of sensor inputs is just as important as refining the model. For example, spectral features tuned to bearing defect frequencies, oil debris or temperature channels, and higher snapshot rates would all lengthen the detectable degradation phase for abrupt failure modes (Randall & Antoni).

The main methodological contributions of this study are the deployment shaped evaluation protocol and the demonstration that this protocol yields an honest, physically interpretable account of both success (precision) and failure (lead time). Future work should aim to enlarge the training dataset with additional run to failure campaigns and evaluate over many more held out trajectories, so that lead time claims carry stronger statistical confidence. Furthermore, future work should aim at enriching the input features to provide more accurate prognostics. Lastly, the pipeline on data needs to be validated from an actual industrial robot, which will close the gap between theoretical testing on the bearing test bench, and the actual robots in question.

Works Cited

Carvalho, Thyago P., et al. “A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance.” Computers & Industrial Engineering, vol. 137, no. 1, Nov. 2019, p. 106024, https://doi.org/10.1016/j.cie.2019.106024. Accessed 13 July 2026.

Chen, Yuanhang, et al. “A Novel Deep Learning Method Based on Attention Mechanism for Bearing Remaining Useful Life Prediction.” Applied Soft Computing, vol. 86, Jan. 2020, p. 105919, https://doi.org/10.1016/j.asoc.2019.105919.

El-Thalji, Idriss, and Erkki Jantunen. “A Summary of Fault Modelling and Predictive Health Monitoring of Rolling Element Bearings.” Mechanical Systems and Signal Processing, vol. 60–61, Academic Press, Mar. 2015, pp. 252–72, https://doi.org/10.1016/j.ymssp.2015.02.008. Accessed 13 July 2026.

Heimes, Felix O. “Recurrent Neural Networks for Remaining Useful Life Estimation.” IEEE Xplore, 1 Oct. 2008, pp. 1–6, https://doi.org/10.1109/PHM.2008.4711422. Accessed 13 July 2026.

Hochreiter, Sepp, and Jürgen Schmidhuber. “Long Short-Term Memory.” Neural Computation, vol. 9, no. 8, Nov. 1997, pp. 1735–80, https://doi.org/10.1162/neco.1997.9.8.1735. Accessed 13 July 2026.

Jardine, Andrew K. S., et al. “A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance.” Mechanical Systems and Signal Processing, vol. 20, no. 7, Oct. 2006, pp. 1483–510, https://doi.org/10.1016/j.ymssp.2005.09.012. Accessed 13 July 2026.

Lei, Yaguo, et al. “Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction.” Mechanical Systems and Signal Processing, vol. 104, May 2018, pp. 799–834, https://doi.org/10.1016/j.ymssp.2017.11.016. Accessed 13 July 2026.

Li, Xiang, et al. “Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks.” Reliability Engineering & System Safety, vol. 172, Apr. 2018, pp. 1–11, https://doi.org/10.1016/j.ress.2017.11.021.

Mobley, R. Keith. An Introduction to Predictive Maintenance. Elsevier, 2002.

Qiu, Hai, et al. “Wavelet Filter-Based Weak Signature Detection Method and Its Application on Rolling Element Bearing Prognostics.” Journal of Sound and Vibration, vol. 289, no. 4–5, Feb. 2006, pp. 1066–90, https://doi.org/10.1016/j.jsv.2005.03.007.

Randall, Robert B., and Jérôme Antoni. “Rolling Element Bearing Diagnostics—A Tutorial.” Mechanical Systems and Signal Processing, vol. 25, no. 2, Feb. 2011, pp. 485–520, https://doi.org/10.1016/j.ymssp.2010.07.017. Accessed 9 July 2026.

Sateesh Babu, Giduthuri, et al. “Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life.” Database Systems for Advanced Applications, 2016, pp. 214–28, https://doi.org/10.1007/978-3-319-32025-0_14.

Saxena, Abhinav, et al. “Damage Propagation Modeling for Aircraft Engine Run-To-Failure Simulation.” 2008 International Conference on Prognostics and Health Management, Oct. 2008, https://doi.org/10.1109/phm.2008.4711414.

Sikorska, J. Z., et al. “Prognostic Modelling Options for Remaining Useful Life Estimation by Industry.” Mechanical Systems and Signal Processing, vol. 25, no. 5, July 2011, pp. 1803–36, https://doi.org/10.1016/j.ymssp.2010.11.018.

Susto, Gian Antonio, et al. “Machine Learning for Predictive Maintenance: A Multiple Classifier Approach.” IEEE Transactions on Industrial Informatics, vol. 11, no. 3, June 2015, pp. 812–20, https://doi.org/10.1109/tii.2014.2349359.

Vaswani, Ashish, et al. “Attention Is All You Need.” Cornell University, 12 June 2017, https://arxiv.org/abs/1706.03762. Accessed 13 July 2026.

Zheng, Shuai, et al. “Long Short-Term Memory Network for Remaining Useful Life Estimation.” 2017 IEEE International Conference on Prognostics and Health Management (Icphm), June 2017, https://doi.org/10.1109/icphm.2017.7998311.

Appendix

Figure 3

Baseline normalized degradation trajectories of the four run-to-failure experiments

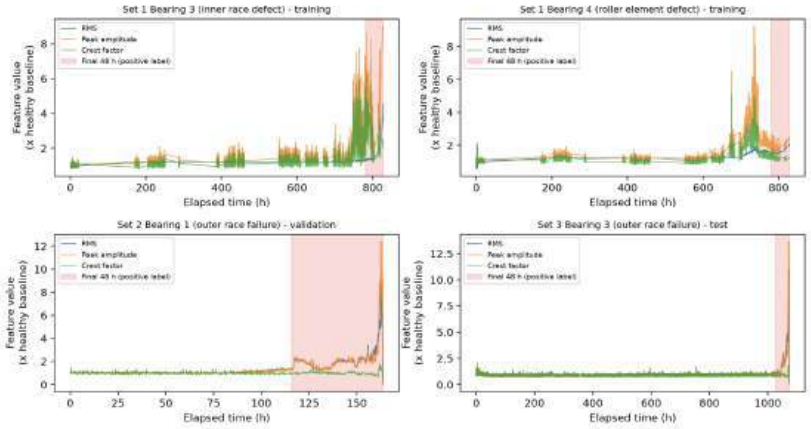


Figure 4

Pearson correlation matrix of input features and labels

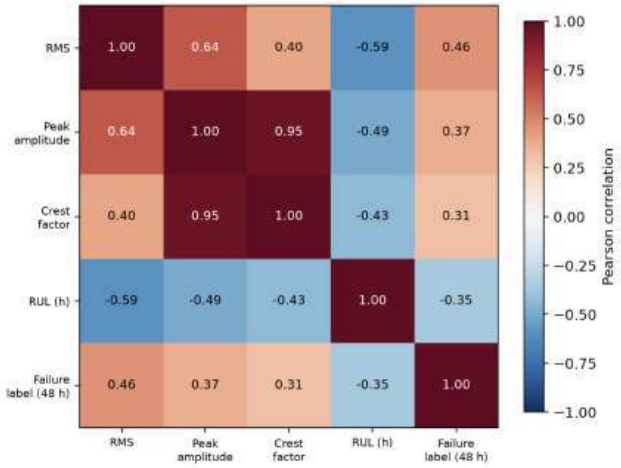


Figure 5

Predicted failure probability on the validation run (Set 2 Bearing 1)

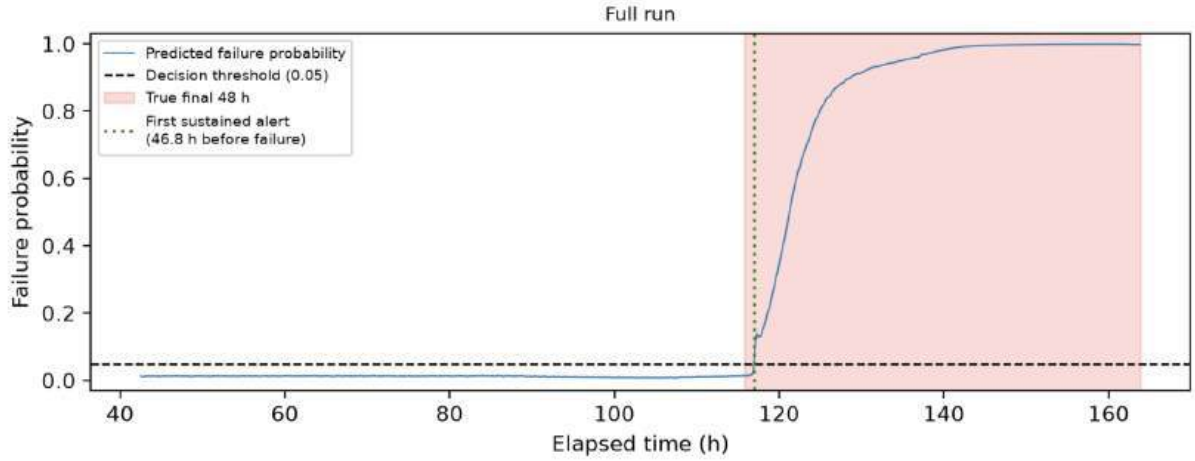


Figure 6

Predicted failure probability on the held-out test run (Set 3 Bearing 3)

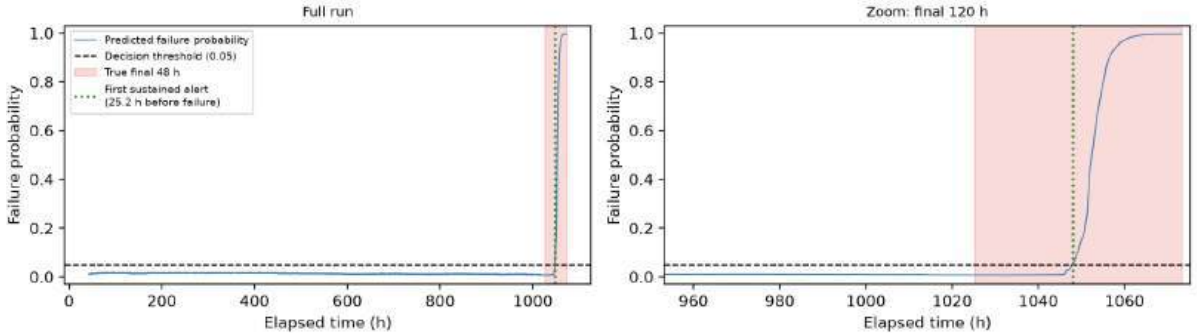


Figure 7

Precision and recall as a function of decision threshold

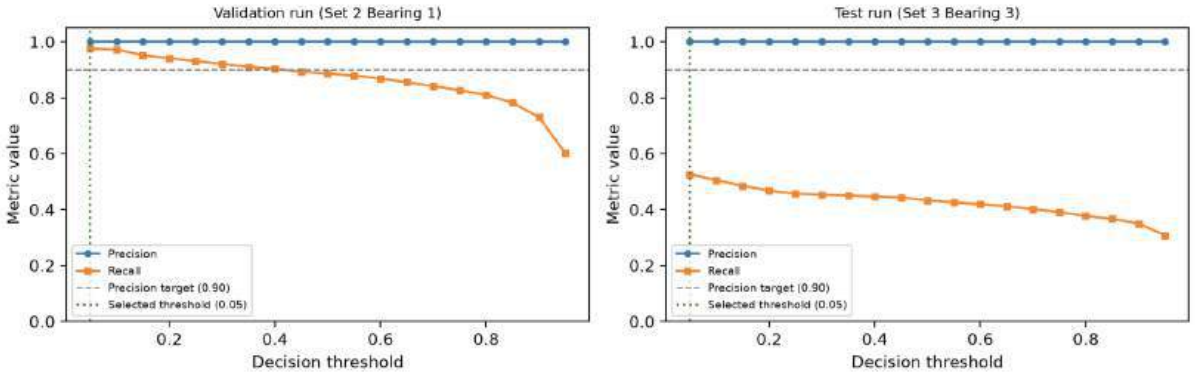


Figure 8

Test confusion matrix at frozen threshold 0.05

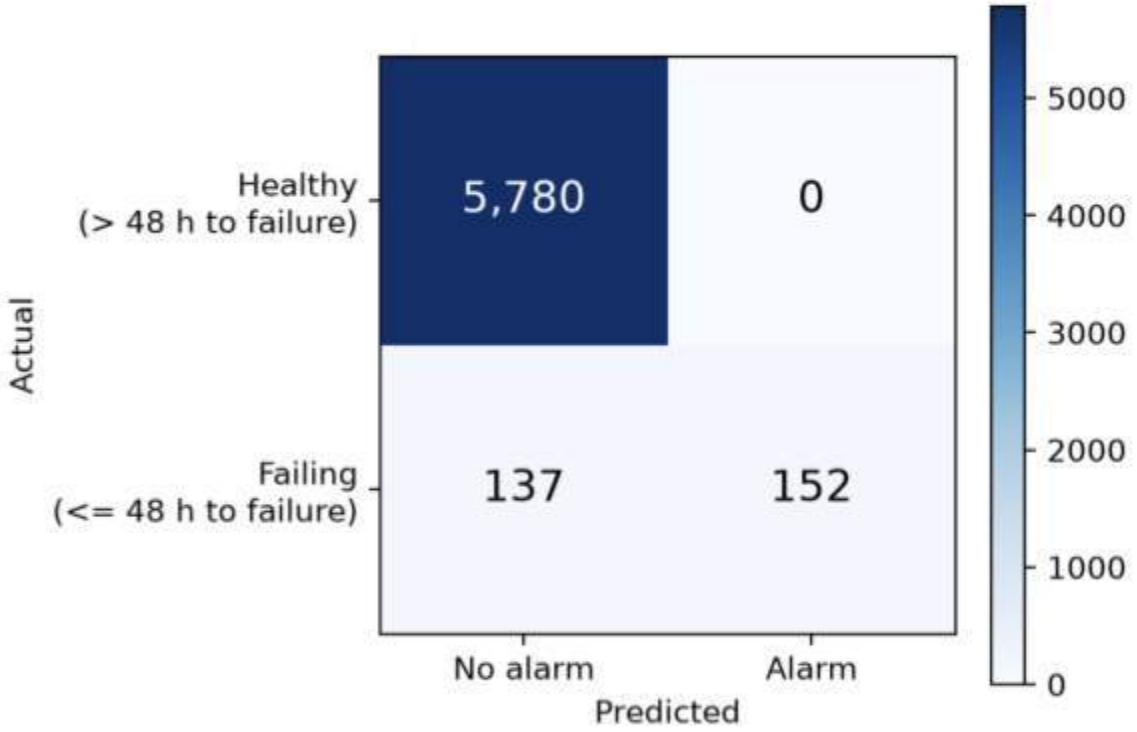
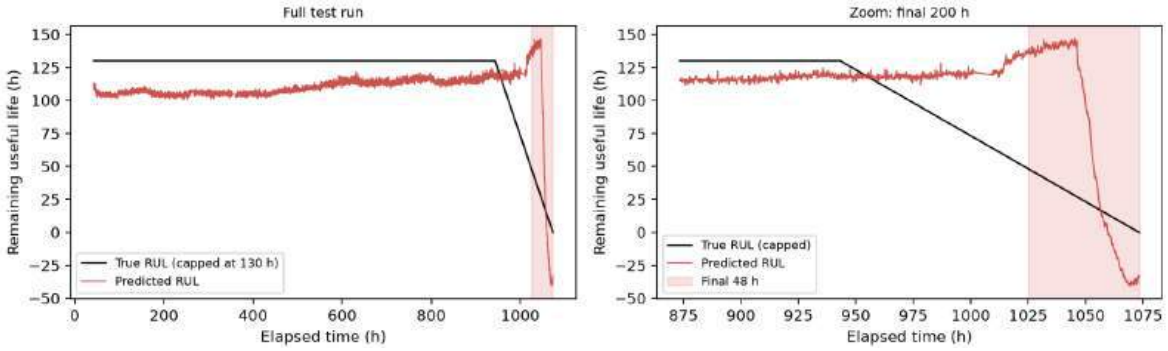


Figure 9

Predicted vs True remaining useful life on the held out test run



Global Monetary and Central Bank Liquidity: Academic Foundations and the Current Global Landscape By Anya Patel, Mentor By Anamika Singh

Abstract

This paper reviews the academic literature on global monetary and central bank liquidity and connects that literature to current conditions as of mid-2026. Global liquidity is understood in the literature as a composite of policy-driven (official) liquidity created by central banks, market-driven (private) liquidity created by cross-border banking and securities activity, and a risk-aversion component reflecting investor sentiment. The paper traces the theoretical foundation of this field, Helene Rey's Global Financial Cycle framework, and its empirical confirmation in Miranda-Agrippino and Rey, before examining the transmission channels through which liquidity moves across borders, including bank balance-sheet constraints, quantitative easing spillovers, and shifts in the composition of cross-border credit since 2008. The paper reviews evidence of regional heterogeneity in how economies absorb liquidity shocks, presents current empirical indicators of global liquidity conditions as of mid-2026, and reports selected quantitative findings from the transmission literature. The synthesis suggests that global liquidity is best understood not as a single undifferentiated tide but as a composite of distinct and at times offsetting forces, whose net effect on any given economy depends on that economy's financial structure, exchange rate regime, and exposure to specific advanced-economy central banks.

1. Introduction

Few concepts in international finance have attracted as much sustained academic attention over the past two decades as global liquidity. The term describes the ease with which capital can move across borders and finance economic activity worldwide, and it has become central to understanding why financial conditions in one country - most often the United States - appear to shape asset prices, credit growth, and capital flows almost everywhere else. The 2008 global financial crisis, the subsequent decade of near-zero interest rates and quantitative easing, the COVID-19 pandemic response, and the 2022-2023 tightening cycle have each, in turn, generated large natural experiments that researchers have used to test and refine theories of how liquidity is created, measured, and transmitted internationally.

This paper has four aims. First, it reviews how the literature defines and decomposes global liquidity into official and private components. Second, it traces the theoretical and empirical development of the Global Financial Cycle framework, which has become the dominant lens through which economists understand cross-border liquidity spillovers. Third, it reports selected quantitative findings from the empirical transmission literature, including magnitude estimates of how liquidity shocks move exchange rates, stock prices, and interest rates. Fourth, it connects this body of research to conditions observed as of July 2026, presenting current empirical indicators of a period marked by an unusual divergence between an easing Federal Reserve and tightening peer central banks.

2. Conceptual Framework: Defining Global Liquidity

2.1 Official and Private Liquidity

Global liquidity is broadly defined as the availability of funds for the purchase of goods, services, and financial assets on a worldwide scale (Eickmeier et al.). Because no single indicator fully captures this condition, the literature typically decomposes global liquidity into three components: policy-driven liquidity, created through the monetary policy actions of advanced-economy central banks; market-driven liquidity, created by banks and non-bank financial institutions through their cross-border lending and securities operations; and a risk-aversion component, reflecting the willingness of global investors to bear financial risk (Le et al.).

Central banks supply official liquidity through conventional tools, principally the setting of short-term policy interest rates, and through unconventional tools such as large-scale asset purchases and emergency lending facilities. Large-scale asset purchase programs expand a central bank's balance sheet directly, injecting a substantial volume of reserves into the banking system (Le et al.). However, official liquidity is only one part of the global picture. Financial institutions independently generate market liquidity through their own cross-border operations, and in an environment of high capital mobility, this private liquidity frequently dominates the liquidity that central banks create directly (Le et al.).

2.2 Measuring Global Liquidity

Eickmeier et al. approach the measurement problem using a global factor model, concluding that fluctuations in worldwide liquidity conditions are driven primarily by three common factors: global monetary policy, global credit supply, and global credit demand. This factor-based approach has proven influential because it avoids relying on any single price or quantity variable - such as a policy rate or a central bank balance sheet figure - that might capture only one dimension of a multidimensional phenomenon. Later empirical work has built on this approach by adding bank-level and country-level granularity, allowing researchers to trace not just whether global liquidity has expanded or contracted, but through which specific channels and toward which specific borrowers that liquidity has flowed (Cerutti et al.).

3. The Global Financial Cycle

3.1 Rey's Dilemma Not Trilemma

The single most influential contribution to this literature is Helene Rey's analysis of what she termed the “Global Financial Cycle” (Rey). Rey demonstrates that capital flows, credit growth, and asset prices across a large number of countries co-move in a common cycle that closely tracks the CBOE Volatility Index (VIX), a widely used proxy for investor risk aversion. Countries that receive heavier credit inflows during periods of low risk aversion are correspondingly more exposed to reversals when risk aversion rises. Critically, this cycle does

not necessarily align with the domestic business cycle of any individual country, meaning that a country's financial conditions can tighten or loosen for reasons that have little to do with its own economic fundamentals.

The policy implication that follows is substantial. Classical open-economy macroeconomics holds that a country can achieve, at most, two of the following three objectives simultaneously: a fixed exchange rate, free capital mobility, and independent monetary policy - a proposition known as the "trilemma." Rey argues that the existence of a global financial cycle collapses this trilemma into a "dilemma": under conditions of free capital mobility, independent monetary policy is achievable only if a country actively manages its capital account in some form, regardless of whether its exchange rate floats or is pegged. A floating exchange rate, in other words, does not by itself insulate a country from the global financial cycle.

Rey further argues that a principal driver of the global cycle is monetary policy in the "center country" of the international financial system, which in practice means the United States. Under this view, U.S. monetary policy shapes the leverage of global financial intermediaries, the volume of cross-border capital flows, and the pace of credit growth throughout the international system, extending the reach of U.S. policy well beyond its own borders.

3.2 Empirical Confirmation: Miranda-Agrippino and Rey

Rey's theoretical claims were subsequently subjected to rigorous empirical testing by Miranda-Agrippino and Rey, who estimate a dynamic factor model on a broad panel of world asset prices and confirm that a single global factor explains a substantial share of the common variation in risky asset prices worldwide ("U.S. Monetary Policy"). Using a Bayesian vector autoregression, the authors show that contractionary U.S. monetary policy shocks produce significant deleveraging among global financial intermediaries, a decline in the supply of credit across countries, a marked retrenchment of international credit flows - particularly through the banking sector - and a broad tightening of financial conditions abroad. Notably, the authors find that countries operating floating exchange rate regimes experience financial spillovers of comparable magnitude to those with fixed regimes, providing direct empirical support for Rey's dilemma-not-trilemma thesis.

In a related and more targeted study, Miranda-Agrippino and Rey examine whether the transmission of U.S. monetary policy to the global financial cycle changed after the 2008 crisis ("Global Financial Cycle after Lehman"). They find that the effects of policy actions targeting short-term interest rates remained broadly similar to the pre-crisis period, but that a distinct "information effect" became more potent at the long end of the yield curve: declines in ten-year U.S. Treasury yields have become associated with weaker global financial activity and flight-to-safety behavior, suggesting the transmission mechanism has evolved even as the underlying cycle persists.

3.3 Critiques and Extensions

Not all subsequent research has found the global financial cycle to be a constant, unchanging force. Some studies find that its grip on capital flows varies meaningfully over time and depends on the specific indicators used to measure it, with several analyses suggesting the cycle's influence weakened somewhat in the years following the 2008 crisis relative to the pre-crisis period. Work on cross-border bank flows similarly finds that global factors explain a meaningful but incomplete share of the variation in international liquidity, leaving substantial room for country-specific and lender-specific characteristics to matter.

4. Transmission Channels of Global Liquidity

Beyond the broad claim that a global financial cycle exists, a large empirical literature - much of it produced by the Federal Reserve Bank of New York, the Bank for International Settlements, and the International Banking Research Network - has sought to identify the precise channels through which liquidity moves across borders.

4.1 The Bank Lending Channel and Balance-Sheet Constraints

Goldberg's review of the drivers of global liquidity emphasizes that spillovers depend heavily on the balance-sheet constraints of the institutions that intermediate capital internationally. Banks with stronger capital positions and greater retained earnings are able to borrow at lower cost, extend more credit, and pass through less of a foreign central bank's policy tightening into their own domestic lending decisions.

Bruno and Shin provide an earlier and complementary account of this mechanism, describing how cross-border banking flows are shaped by the leverage cycles of global banks. In their framework, an increase in local currency appreciation raises the capacity of local borrowers to take on foreign-currency debt, which in turn draws in additional cross-border bank lending; the resulting expansion of bank balance sheets amplifies the initial liquidity shock rather than simply transmitting it in a linear fashion. This amplification mechanism has become a standard building block in subsequent models of global liquidity transmission.

4.2 Quantitative Easing Spillovers

A widely cited study by Chen et al., first circulated by the Bank for International Settlements in 2012 and later republished in the Journal of International Money and Finance, examines the cross-border effects of quantitative easing programs conducted by major advanced-economy central banks. The authors find clear evidence that QE announcements moved emerging-market asset prices broadly: equity valuations rose, government and corporate bond yields fell, and credit-default-swap spreads narrowed. They attribute these effects primarily to shifts in market expectations about future cross-border capital flows and the anticipated policy responses of emerging-market central banks.

Related work on international spillovers of quantitative easing finds an important asymmetry: conventional monetary easing in a large economy tends to raise output in other countries, whereas quantitative easing tends to have the opposite effect, undermining the

international competitiveness of trading partners and depressing their aggregate output. This asymmetry suggests that the specific policy instrument used to ease financial conditions carries meaningfully different implications for the rest of the world.

4.3 Cross-Border Banking and the Shifting Composition of Global Liquidity

Avdjiev et al. document a structural shift in how global liquidity has been delivered across borders since the 2008 financial crisis. Analyzing data across a large panel of countries, they find a reallocation away from traditional cross-border bank lending and toward international debt securities markets as the dominant channel of global liquidity provision. Their analysis further shows that liquidity flows become markedly more sensitive to risk conditions when the banks or non-bank intermediaries that channel them face tighter balance-sheet constraints, reinforcing the balance-sheet mechanisms identified by Goldberg and Bruno and Shin.

4.4 Push and Pull Factors in Capital Flows

A parallel strand of literature frames cross-border liquidity transmission in terms of "push" and "pull" factors driving capital flows to emerging markets. Push factors are external conditions set in advanced economies - global interest rates, risk aversion, and the overall stance of monetary policy - while pull factors are domestic characteristics of the recipient country, including growth prospects, institutional quality, and the openness of its capital account (Fratzscher). Fratzscher's analysis of the 2008 crisis finds that push factors, particularly US monetary policy and global risk conditions, were the dominant determinants of capital flow behavior during that episode.

Cerutti et al. refine this picture considerably by examining gross capital inflows to a panel of thirty-four emerging markets. They find that aggregate inflows co-move substantially across countries, but that this co-movement is concentrated in bank-related and portfolio flows rather than foreign direct investment. Perhaps most notably, they find that a country's sensitivity to global push factors depends less on its own institutional fundamentals than on the composition of its foreign investor base and the structural characteristics of its local financial markets.

5. Regional and Country-Level Heterogeneity

5.1 Divergent Responses in the Asia-Pacific

While much of the literature reviewed above treats advanced-economy monetary policy as a largely undifferentiated source of global liquidity, more recent research has emphasized that different central banks transmit liquidity shocks through markedly different channels, with correspondingly different effects on recipient economies. Le et al. study liquidity spillovers to eight Asia-Pacific economies - Hong Kong, Indonesia, Korea, Malaysia, the Philippines, Singapore, Thailand, and Vietnam - over the period 2000 to 2022, using a panel vector autoregression model that distinguishes between policy-driven and market-driven liquidity shocks.

Their central finding runs counter to much of the earlier literature: Asia-Pacific economies are more sensitive, on balance, to market-driven liquidity - cross-border bank credit flows, particularly those denominated in U.S. dollars - than to policy-driven liquidity originating from advanced-economy central banks. By contrast, shocks originating from the Bank of Japan - both its shadow short rate and its balance sheet - measurably depress inflation and generate temporary depreciation pressure on Asian currencies, reflecting the historically close financial and trade linkages between Japan and the rest of the region.

5.2 Quantitative Evidence from Impulse-Response Estimates

Table 1 summarizes selected magnitude estimates reported by Le et al. from impulse-response functions estimated over a ten-quarter horizon. These figures illustrate the differing scale and persistence of policy-driven versus market-driven liquidity shocks on Asia-Pacific macro-financial variables.

Shock Type	Variable Affected	Estimated Impact	Persistence
Market-driven liquidity (USD credit flows)	Stock prices	+0.5 pct. pts. or more on impact	Short-run (1 period)
Market-driven liquidity (USD credit flows)	Exchange rate (FOREX)	-0.2 pct. pts. or more (appreciation)	Immediate
Fed shadow-rate / balance-sheet shock	Stock prices	~2 pct. pt. jump, then sharp reversal	1 period, reverses quickly
BoJ shadow-rate shock	Inflation (CPI)	Negative effect	Persists several periods
BoJ balance-sheet shock	Exchange rate (FOREX)	Temporary depreciation pressure	Short-run

Table 1. Selected impulse-response estimates of global liquidity shock transmission to Asia-Pacific economies. Source: Le et al., Empirical Economics, vol. 67, no. 3, 2024, pp. 1091-1113.

5.3 Inflation Targeting and Policy Buffers

A separate strand of the literature examines whether domestic policy frameworks can moderate an economy's exposure to global liquidity shocks. Several studies find that emerging markets operating under inflation-targeting regimes experience less pronounced volatility in output, inflation, and asset prices in response to global liquidity shocks than emerging markets operating under alternative monetary frameworks. This finding suggests that, while no domestic policy framework can fully insulate a country from the global financial cycle - consistent with Rey's dilemma thesis - certain frameworks can meaningfully reduce the magnitude of the resulting volatility.

6. Current Global Liquidity Conditions: A Mid-2026 Snapshot

The theoretical and empirical literature reviewed above maps closely onto conditions observed as of July 2026. Table 2 reports selected quantitative indicators of the current global liquidity environment, compiled from financial data providers and market commentary current as of July 2026.

Indicator	Value	Reference Date
U.S. M2 money supply	≈ $22.65 trillion	February 2026
U.S. M2 money supply (prior month)	≈ $22.44 trillion	January 2026
Fed Reserve Management Purchases (T-bills)	≈ $40 billion / month	Since January 2026
Projected ECB + BoE + BoJ balance-sheet reduction	≈ $1.2 trillion	Full-year 2026 (projected)
Fed overnight reverse repo (ON RRP) usage	≈ $26.9 billion	June 30, 2026
Fed ON RRP peak usage (post-2020)	> $2.5 trillion	2021-2022 peak
G4 central banks' combined pandemic-era asset purchases	≈ $7.7 trillion	2020-2021
G4 central banks' projected cumulative unwind	≈ $5 trillion (~two-thirds reversed)	By end of 2026 (projected)

Table 2. Selected global liquidity indicators, mid-2026. Figures compiled from Trading Economics, MacroMicro, the Mises Institute, and market-strategist summaries of G4 central bank balance-sheet projections; provided for illustrative, current-conditions context rather than as peer-reviewed academic findings.

The U.S. Federal Reserve ended the passive runoff of its balance sheet in December 2025 and has since conducted the Reserve Management Purchases program shown in Table 2, aiming to maintain an ample level of bank reserves alongside a series of policy rate cuts. U.S. M2 money supply has grown in tandem with this shift, reaching approximately $22.65 trillion by February 2026 - among the fastest pace of M2 growth in several years.

At the same time, the European Central Bank, the Bank of England, and the Bank of Japan are together projected to reduce their combined balance sheets by approximately $1.2 trillion over the course of 2026. Some market analysts argue that continued tightening among the Federal Reserve's peer institutions could largely offset the effect of the Fed's own renewed expansion at the aggregate global level. This divergence illustrates the point, made by both Rey and Miranda-Agrippino and Rey, that the center country's monetary stance interacts with, rather than dictates in isolation, the policies of other major central banks in shaping global financial conditions.

A further development concerns the depletion of the balance-sheet buffers that absorbed liquidity strain during the post-2020 period of monetary expansion. As Table 2 shows, usage of the Federal Reserve's overnight reverse repurchase agreement facility fell from a peak above $2.5 trillion to approximately $26.9 billion by the end of June 2026. With this buffer largely

exhausted, new U.S. government debt issuance increasingly competes directly for bank reserves, deposits, and dealer balance-sheet capacity - a more fragile configuration of the short-term funding system than existed only a few years earlier, and one consistent with the balance-sheet-constraint mechanisms identified by Goldberg and by Avdjiev et al.

Viewed across the full post-pandemic cycle, the four major central banks discussed throughout this paper collectively purchased approximately $7.7 trillion in assets during 2020 and 2021. Current projections suggest that, by the end of 2026, these institutions will have collectively unwound roughly $5 trillion of that expansion - reversing approximately two-thirds of the pandemic-era stimulus even as the Federal Reserve itself has paused and partially reversed its own tightening. This trajectory places 2026 in an unusual intermediate position: neither a return to the ultra-accommodative conditions of 2021 nor a continuation of the aggressive tightening seen in 2022 and 2023, but rather a genuine divergence across the world's most influential monetary authorities.

7. Discussion: Implications for Policy and Markets

The literature reviewed in this paper converges on several implications relevant to both policymakers and market participants. First, because global liquidity is jointly produced by official and private sources that interact dynamically rather than operating independently (Le et al.), policy analysis that focuses solely on central bank balance sheets while ignoring cross-border bank and securities market activity is likely to produce an incomplete picture of prevailing financial conditions. Second, the balance-sheet health of the specific institutions that intermediate capital - a theme running through Bruno and Shin, Goldberg, and Avdjiev et al. - means that the transmission of any given liquidity shock depends as much on the condition of the financial system receiving it as on the magnitude of the shock itself. Third, and perhaps most practically, the finding that different advanced-economy central banks transmit shocks through different channels (Le et al.) - illustrated quantitatively in Table 1 - implies that a one-size-fits-all model of “the” global financial cycle risks obscuring economically meaningful regional variation.

For market participants, the practical takeaway from the current 2026 snapshot in Table 2 is that aggregate global liquidity conditions cannot be inferred from any single central bank's policy stance in isolation. A Federal Reserve that is easing does not necessarily imply that global financial conditions overall are loosening, if peer central banks are tightening by a comparable magnitude at the same time. Similarly, the depletion of the reverse repo buffer illustrates that liquidity “plumbing” can become more fragile even in the absence of any single dramatic policy shift, simply as a byproduct of the cumulative unwind of a multi-year liquidity cycle.

8. Limitations and Directions for Future Research

This paper is a literature synthesis rather than an original empirical study, and it carries the limitations inherent to that approach. First, the review draws primarily on a curated set of influential papers rather than an exhaustive systematic search of the field, so some relevant

contributions - particularly non-English-language research and unpublished working papers - are likely underrepresented. Second, the mid-2026 figures presented in Table 2 are drawn from financial data providers and market commentary rather than from the peer-reviewed literature, and should be treated as illustrative context rather than as findings with the same evidentiary standing as the academic sources reviewed elsewhere in this paper.

Several directions for future research follow naturally from the gaps identified above. First, given that Le et al. find Asia-Pacific economies respond more strongly to Bank of Japan policy than to Federal Reserve policy, a natural extension would be to test whether this pattern holds for other regions with strong trade and financial ties to Japan, such as parts of Latin America or the Middle East, or whether it is a distinctly Asia-Pacific phenomenon tied to regional financial integration. Second, because Avdjiev et al. document a shift from bank lending toward debt securities as the dominant channel of global liquidity, future work could examine whether this shift has continued or reversed in the aftermath of the 2022-2023 tightening cycle, when bond market volatility rose sharply. Third, the current episode of Federal Reserve easing alongside continued tightening by other major central banks - documented in Table 2 - is itself a natural experiment that future researchers could use to test whether Rey's global financial cycle framework predicts a net loosening, a net tightening, or a genuinely ambiguous outcome for emerging-market financial conditions, once the relevant data are available with the benefit of hindsight.

9. Conclusion

This paper has reviewed the academic literature on global monetary and central bank liquidity, tracing its development from foundational definitional work through the Global Financial Cycle framework and into a growing body of research on regional heterogeneity in liquidity transmission, supported by quantitative evidence drawn directly from the empirical literature and current market data. Several conclusions emerge clearly: global liquidity is a composite phenomenon rather than a single variable; its transmission depends heavily on the balance-sheet condition of intermediating institutions; and different advanced-economy central banks propagate liquidity shocks through distinct channels that affect different regions unevenly, as the magnitude estimates in Table 1 illustrate. Conditions observed as of mid-2026 and reported in Table 2 - an easing Federal Reserve, tightening peer central banks, and a markedly thinner buffer in short-term funding markets - offer a timely and concrete illustration of these theoretical mechanisms in practice, underscoring the continued relevance of this literature for understanding the international financial system as it evolves.

Works Cited

Avdjiev, Stefan, et al. “The Shifting Drivers of Global Liquidity.” Journal of International Economics, vol. 125, 2020, article 103324, doi:10.1016/j.jinteco.2020.103324.

Bruno, Valentina, and Hyun Song Shin. “Cross-Border Banking and Global Liquidity.” Review of Economic Studies, vol. 82, no. 2, 2015, pp. 535-64, doi:10.1093/restud/rdu042.

Cerutti, Eugenio, et al. “Push Factors and Capital Flows to Emerging Markets: Why Knowing Your Lender Matters More than Fundamentals.” Journal of International Economics, vol. 119, 2019, pp. 133-49, doi:10.1016/j.jinteco.2019.04.006.

Chen, Qianying, et al. “International Spillovers of Central Bank Balance Sheet Policies.” BIS Papers, no. 66, Bank for International Settlements, Oct. 2012, pp. 230-64, www.bis.org/publ/bppdf/bispap66p.pdf. Republished in Journal of International Money and Finance, vol. 67, 2016, pp. 62-81.

Eickmeier, Sandra, et al. “Understanding Global Liquidity.” European Economic Review, vol. 68, 2014, pp. 1-18, doi:10.1016/j.euroecorev.2014.01.015.

Fratzscher, Marcel. “Capital Flows, Push versus Pull Factors and the Global Financial Crisis.” Journal of International Economics, vol. 88, no. 2, 2012, pp. 341-56, doi:10.1016/j.jinteco.2012.05.003.

Goldberg, Linda S. Global Liquidity: Drivers, Volatility and Toolkits. Staff Report no. 1064, Federal Reserve Bank of New York, 2023, www.newyorkfed.org/medialibrary/media/research/staff_reports/sr1064.pdf.

Le, Chau, et al. “Global Liquidity Spillovers in the Asia-Pacific Region: Policy-Driven versus Market-Driven Effects.” Empirical Economics, vol. 67, no. 3, 2024, pp. 1091-1113, doi:10.1007/s00181-024-02573-4.

Miranda-Agrippino, Silvia, and Helene Rey. “The Global Financial Cycle after Lehman.” AEA Papers and Proceedings, vol. 110, 2020, pp. 523-28, doi:10.1257/pandp.20201096.

---. “U.S. Monetary Policy and the Global Financial Cycle.” Review of Economic Studies, vol. 87, no. 6, 2020, pp. 2754-76, doi:10.1093/restud/rdaa019.

Rey, Helene. Dilemma Not Trilemma: The Global Financial Cycle and Monetary Policy Independence. NBER Working Paper no. 21162, National Bureau of Economic Research, 2015, doi:10.3386/w21162.

Acknowledgments

The author thanks Anamika Singh for her guidance and mentorship throughout the research and writing of this paper.

Data Note

U.S. M2 money supply figures in Table 2 are drawn from the Federal Reserve Board's H.6 Money Stock Measures release and the FRED database maintained by the Federal Reserve Bank of St. Louis, the primary official sources for this series. Figures describing Federal Reserve Reserve Management Purchases, projected 2026 balance-sheet reductions at the ECB, Bank of

England, and Bank of Japan, and overnight reverse repo facility usage were supplemented with financial data providers and market commentary current as of July 2026, including Trading Economics, MacroMicro, and market-strategist summaries of G4 central bank balance-sheet projections, since real-time figures for these specific series are not always available in a single official release at the time of writing. These supplementary figures are provided for illustrative, current-conditions context rather than as peer-reviewed academic findings, and should be independently verified against primary central bank releases before use in a formal submission.

Circular Business Model Innovation and the Decoupling Question: A Case Study of H&M Group By Sravya Tatiraju

Abstract

It is estimated that the fashion sector constitutes approximately 10% of the annual global greenhouse gas emissions, which substantiates the move from the linear production model of ‘take-make-dispose’ to that of a circular business model. The given paper presents a secondary qualitative case study of the Circular Business Model Innovation (CBMI) of H&M Group, with the objective to find out to what degree the circular initiatives of the company lead to the reduction of its absolute impact and not only its impact per unit of output. The effectiveness is estimated along three dimensions: (1) absolute versus relative decoupling of the emissions from the production growth; (2) the level of closed circularity in terms of material recovery as compared to open circularity of substitution; (3) the level of customer engagement in the circular services as compared to corporate ambition. The research relies on the data from three secondary resources: H&M's own sustainability disclosures for the year 2024, tested literature on circular economy and life cycle assessment, and coded excerpts from two qualitative studies of consumer attitudes towards sustainability applicable to H&M. However, as the latter studies incorporated the opinions of respondents concerning the principles of sustainability used by fast fashion producers in general and not concerning the CBMI in particular, the findings achieved are regarded as an indicator rather than a basis for direct evidence. The investigation reveals that the company is effective in terms of relative decoupling, but not in terms of the absolute decoupling, since it fails to satisfy its own expectations regarding emissions reduction, as the Scope 3 emissions kept rising steadily while the percentage of recycled and sustainably produced materials kept increasing. It is observed that consumer-related hurdles (such as the failure to gain trust, sensitivity to prices, and inadequate knowledge about the manufacturing process) and structural challenges (such as the incongruence between the production level and the closed-loop recycling) appear to be interactive in nature. The paper concludes that CBMI can be employed effectively as a regulatory method, managing to act at an effective level, nevertheless, it won’t help to achieve the goal of turning the CBMI into a perfect method of reducing environmental impact. This leads to further investigation into those areas involving the development of governance and business models.

Keywords

Social and Behavioral Sciences; Economics; Circular Economy; Fast Fashion; Business Model Innovation; Decoupling

Introduction

The global fast fashion market has a value of $2.4 trillion; however, this area is known as one of the most influential polluters, comprising an average of 10% of global carbon emissions and 20% of water pollution caused by industry (Niinimäki et al.). The traditional model of fast

fashion mainly follows the pattern of 'take-make-dispose,' which presupposes the acquisition of resources (cotton, crude oil, and polyester) for garments that would be likely demanded and disposed of shortly after such a purchase. However, as the raw materials become scarce and the costs of employing this business model become unbearable, the need for such an approach diminishes significantly.

As an answer to the given issue, businesses turn their attention to Circular Business Model Innovation (CBMI), in order to combine economic efficacy with environmental sustainability (Geissdoerfer et al.). CBMI can be defined as the process of rethinking the value chain: how a company's value is created, delivered, and captured. H&M Group, which is among the top fashion retailers globally, serves as an exemplary case illustrating the shift toward circular business models. As far as adoption of circularity-based model innovation (CBMI) is concerned, for H&M, this is more than merely an altruistic initiative; it is also a strategic reaction to changing E.U. regulations related to textile waste management facilitated by the Digital Product Passport (DPP) that will oblige companies to share validated details pertaining to the performance of textiles made, their recyclability, and other important indicators (European Commission). Thus, reaching true circularity means meeting licensing requirements rather than just marketing needs in the case of H&M. In addition to that, CBMI is gaining in importance on a competitive level due to the growing pressure from cost-effective rivals such as Shein and Temu who take away relevant market share.

Although Circular Business Model Innovation (CBMI) has become one of the dominant approaches to corporate sustainability, existing research has primarily examined its implementation and operational efficiency rather than whether these initiatives produce absolute reductions in environmental impact. This study addresses this issue through a qualitative secondary case study of H&M Group. While primary emission and consumer data are out of the scope of this work, the term "effectiveness" in the text does not refer to any exact determination of effectiveness in the study focused on H&M; instead, it addresses the narrower question: taking into consideration H&M’s disclosures and independent life cycle analysis (LCA) and relevant decoupling literature, what information does the qualitative evidence imply about H&M's CBMI being effective in terms of cutting down the company’s environmental influence? Thus, the research will address the questions of (1) whether or not H&M emissions are decoupling from production increase in absolute or relative terms; (2) whether H&M's material-recovery systems are progressing from open-loop substitution, such as recycled polyester from plastic bottles, toward closed-loop, textile-to-textile recycling; and (3) whether consumers are engaging with H&M's circular services at a level consistent with the company's stated ambitions. H&M's own sustainability reporting shows progress on some of these dimensions, particularly the share of recycled content in its garments; independent literature suggests this progress may be offset by continued growth in total production.

This paper argues that H&M has made measurable improvements to its supply chain (value creation) but has made comparatively little progress on how customers interact with its circular services (value delivery) or on aligning corporate governance with circularity outcomes

(value capture). It examines these three levels — corporate and system structure, technology and materials, and consumer behavior — as interacting parts of a single system rather than as independent, freestanding problems.

Literature Review

The Environmental Cost of the Linear Model

The current literature has extensively examined the issues of the unsustainability of the linear model in the fashion industry. According to Niinimäki et al., this is due to a paradox regarding fast fashion: as cheap prices depend on the high volume of sales, brands are forced to manufacture a large number of clothes, ultimately resulting in overproduction. According to the authors, this creates the so-called ecological debt since as the linear model includes a take-make-waste flow, the value generated by clothing is basically destroyed once the clothing becomes outdated. Thus as the scarcity of resources increases, this model is exposed to an increasing level of threat. On the other hand, it forces companies to look for new methods of operation.

Circular Business Model Innovation as a Driver of Change

CBMI differs from the usual product innovation. product innovation may mean developing a new textile, while CBMI deals with reorganizing the process of creation and delivery of the product. Kirchherr et al. define the circular economy (CE) model as an economic activity aimed at decoupling the economy from the use of limited resources by switching the concept of "end-of-life" into "reduction," "reuse," and "recovery." According to Guldmann and Huulgaard, the main obstacles to implementing CBMI are not technical but organizational and behavioral in nature. In order to effectively implement this, companies are required to change their business model.

Empirical evidence on whether this redesign actually improves environmental or business outcomes is more limited, but a small and growing body of firm-level research offers partial support. Motke and others used survey data from industrial chemical companies in Brazil and stated that companies that adopted greater circularity in their business models recorded enhanced market, production, and social success; however, it is important to note that this research used self-reported business performance, rather than measuring impact on the environment. In the research by Ostermann et al. specializing in the fashion industry, it became evident that most of the ten new companies of the circular fashion industry focused mainly on the closed-loop innovation of business models (value creation is based on material recycling and redesign), while the innovations based on the service model of circular fashion business (the business produces items for rent and resale) are often missing. One can see how H&M is struggling to transition to more service-oriented circular business models in the course of their operations. At the same time, the research by Conca and the team, studying over 1000 corporations dealing with food and health, proved that using the principles of circular economy positively affected the financial

standing of companies, which leaves one to conclude that there is no contradiction between applying circularity and earning profits. Nonetheless, none of these works answer the question of whether CBMI allows companies to reduce their total environmental impact, which is the primary goal of this study, in the case of H&M specifically.

The Gap Between Corporate Rhetoric and Consumer Reality

It should be noted that much of contemporary academic discussion related to the sustainable fashion industry revolves around the issue of the "attitude-behavior gap': the gap between consumers' intentions and their actual purchasing behavior. Despite reporting high environmental concern, customers do not consistently translate this sentiment into their purchasing behavior. One of the contributors to this phenomenon is distrust in corporate claims regarding sustainability, also known as greenwashing. The lack of reliability in a company's stated marketing claims, such as a brand promoting sustainability while releasing hundreds of new styles each week, may reduce trust and discourage participation in circular services such as garment take-back programs.

Corporate Production Scale as a Structural Impediment

It is a repeated shortcoming of the current research on CBMI that non-successes in sustainability performances are deemed to be an issue mainly related to operations or consumer perspectives, thus overlooking the importance of corporate governance and scale of production. Thus, it is widely known that H&M Group's total garment production amounts to approximately 3 billion units. Since this statistic comes from journalists and advocacy sources, it is referred to as an approximate external estimate rather than a corporation KPI. The exact amount of production is of little importance because production volumes dictate the efficiency of material recovery: so, improvement of H&M's environmental impact by a unit of garment does not yield the expected result at the aggregate level. This situation is analogous to findings in decoupling research in other areas of the textile industry. Analyzing thirty Chinese provinces over two decades, Hong et al. found that during most periods, carbon emissions in textile production experienced only weak decoupling from economic growth. In other words, while energy efficiency improved, the sheer scale of production remained the key factor driving emissions, a pattern of relative rather than absolute decoupling aligning closely with H&M's own disclosures discussed later in this paper. In contrast, the Circularity Metrics Lab of the European Environment Agency reported at the level of the EU that from 2010 to 2022, greenhouse gas emissions from textile consumption in the EU decreased by 22% per capita despite the consumption growing by 15%. This represents an example of absolute decoupling occurring thanks to the reductions in emission intensity (European Environment Agency). This shows that absolute decoupling in textiles can be achieved in certain conditions, even though H&M has not achieved it in its recent development yet.

Guldmann and Huulgaard argue that institutional obstacles, such as management motivation being connected to revenue growth rather than material recovery, are among the

major obstacles to the transition to a circular economy. Because the growth of production and minimization of material waste act in opposite directions, the company cannot have both maximum output of new garments and minimum material waste; these are competing objectives, and not complementary ones. As a result, the gap in H&M's environmental performance cannot be explained only by the consumers' behavior. H&M is responsible for its production quantity, decides whether to offer recycling incentives together with loyalty programs, and determines the mix of materials and its communication about sustainability. Thus, putting the blame for lack of circularity only on the consumers may lead to missing the establishment of supply chain decisions on the corporate level. However, as the Discussion considers further below, this does not mean consumers bear no responsibility at all.

Analytical Propositions

Because this paper conducts a qualitative secondary-data case study, it does not test statistical hypotheses in the conventional sense. Following case-study methodology, however, it is useful to state in advance the propositions that circular economy and decoupling theory would predict, against which the evidence below is compared.

P1: If CBMI is concentrated in value creation (material substitution) without comparable investment in value delivery (consumer-facing services) or value capture (revenue-model and governance change), the expected outcome is relative decoupling — declining impact per unit — without absolute decoupling, or declining total impact.

P2: Because material substitution requires coordination primarily with suppliers, while service-based and governance change requires broader stakeholder alignment (Guldmann and Huulgaard), value-creation innovation is expected to outpace value-delivery and value-capture innovation within the same firm.

P3: Consumer sentiment recorded in general fast-fashion sustainability research is expected to correlate with, but not directly measure, sentiment about CBMI specifically, since constructs such as greenwashing perception and trust are conceptually adjacent to, but not identical with, circularity-specific attitudes.

These propositions structure the Results and Discussion sections below.

Methods

This paper presents an empirical secondary qualitative case study. Due to the scope of the paper, it does not collect new primary data. Instead, it focuses on the qualitative analysis of existing secondary sources using a structured qualitative coding process, following a case-study format in which a single information-rich case (H&M) is examined against the analytical propositions stated above.

The analysis draws on three secondary sources: (1) H&M's own 2024 Annual and Sustainability Report and related corporate disclosures, used to identify the company's stated objectives and reported key performance indicators; (2) peer-reviewed literature on circular economy, decoupling, and life-cycle assessment, used to benchmark H&M's disclosed figures

against independent industry evidence; and (3) coded excerpts from two existing qualitative studies of consumer attitudes toward H&M's sustainability communication — Ikonomou's master's thesis on the attitude-behavior gap among H&M customers in Denmark, and Zaidi's master's thesis analyzing focus-group attitudes toward H&M's sustainability communication in Sweden.

It must be clarified that neither Ikonomou nor Zaidi asked participants specifically about H&M's Circular Business Model Innovation. Their studies were broader in terms of consumer perception of sustainability in fast fashion and H&M's sustainability communication through semi-structured interviews and focus groups (Ikonomou: six participants, ages 24–29, recruited by snowball sampling in Denmark) and focus groups (Zaidi: twenty participants in Sweden). To maintain an accurate analysis, this paper does not treat these excerpts as direct evidence of consumer responses to CBMI; instead, in line with Proposition 3, the excerpts are used as evidence of general consumer sentiment toward H&M's sustainability communication, such as skepticism and price sensitivity.

The data was coded in a hierarchical fashion whereby the first layer of coding followed the three pillars of business model innovation, namely: value creation (how the good is produced), value capture (how the company generates revenue), and value delivery (how the products and services produced reach the consumer). The second layer of coding is of deductive nature and is embedded solely in the interview data and sorts the excerpts into two codes: Corporate Intent and Consumer Perception. The process of open coding of the interview data has led to the development of three major themes — skepticism, economic friction, and awareness, that will be discussed in the Results section below. After that, H&M's KPIs were compared to qualitative analysis and data derived from independent sources to find out whether the findings provided by H&M are valid based on the information available.

Results

Corporate KPIs and the Decoupling Question

H&M Group's latest sustainability KPIs are presented in Table 1. In 2024, H&M claims that 89% of its products come from recycled or "sustainably sourced" materials. This category, according to the company's disclosure, incorporates organic, recycled, and regeneratively grown fibers as well as materials certified according to sustainability standards. This data suggests an upward revision of the previously quoted numbers, which were around 85%. The company accounts for recycled materials and indicates an increase from approximately 26% to 29% as well as growth for recycled polyester from 79% to 94%.

And emissions data tells a different story that highlights the differentiation between relative and absolute decoupling concepts. The reported Scope 1 and 2 emissions (operational emissions and emissions associated with electricity and heating deliveries) dropped by 41% as compared with 2019 baselines, while the Scope 3 emissions (indirect emissions appearing everywhere along the supply chain) experienced a 23.7% drop. Nevertheless, between years,

Scope 3 emissions slightly increased by 3% to 3.5% between 2023 and 2024, according to the company and expert assessments caused by air cargo deliveries and heavier transported materials. Furthermore, the company announced its total weight (524,739 metric tons) which remained practically the same as in 2023 and is less than results reported in 2022 (561,087 metric tons); thus, the increase in Scope 3 emissions was most likely not connected with the increasing amounts of produced products.

The observed trend is in agreement with Proposition 1, whereby H&M, in the long-term period from 2019 until 2024 reveals significant relative decoupling which features some absolute decoupling in Scope 1 and 2 GHG emissions, yet has a short-term period (just from 2023 up to 2024) suggesting that the negative trend in Scope 3 emissions exists despite the progress of H&M towards increasing the share of recycled materials. Since Scope 3 constitutes a major portion of H&M's total footprint, this incident is highly informative in terms of absolute impact for H&M which exceeds what recycling percentage shows. For the sake of comparison, Hong and his colleagues showed in their study of the Chinese textile sector that the total decoupling achieved was not really strong owing to the continued growth of production amount, although H&M shows a positive long-term trend with much better results similar to the EU showing absolute decoupling for textile consumption in general.

Table 1. H&M Group Selected Sustainability KPIs, 2022–2024 (H&M Group; secondary trade coverage synthesizing that report)

Indicator	**2022**	**2023**	**2024**	**Notes**
Recycled or sustainably sourced materials (share of total)	—	83%	89%	Corrects the 85% figure sometimes cited in earlier commentary
Recycled materials specifically (share of total)	—	~26%	29.5%	Target: 30% by 2025; 50% by 2030
Recycled polyester (share of polyester used)	—	79%	94%	—
Scope 1 + 2 emissions vs. 2019 baseline	~−24%	—	−41%	Absolute decoupling on this measure
Scope 3 emissions vs. 2019 baseline	—	—	−23.7%	Relative decoupling over the longer run
Scope 3 emissions, year-over-year	—	—	+3 to 3.5% vs. 2023	Short-run reversal despite materials progress
Disclosed total product weight (metric tons)	561,087	not separately identified in sources reviewed	524,739	H&M's own disclosed production-weight metric
Externally estimated annual garment production	≈ 3 billion units/year (recurring media estimate)	—	—	Not an H&M-disclosed KPI; included for context only

Findings from Consumer-Facing Secondary Data

Tables 2 and 3 reproduce coded excerpts from Ikonomou's and Zaidi's studies, organized around the three themes identified through open coding: skepticism and trust, economic and behavioral barriers, and operational friction and awareness. As discussed in Methods, none of these excerpts refer to H&M's CBMI by name; they reflect participants' general views of H&M's sustainability communication and of fast-fashion sustainability more broadly. They are presented here as coded secondary data, and their cautious implications for CBMI specifically are developed in the Discussion section that follows.

Table 2. Coded excerpts on general consumer sustainability attitudes toward H&M, drawn from six semi-structured interviews conducted in Denmark (Ikonomou**)**

Theme & Codes	**Representative Participant Excerpts**
Skepticism & Trust Deficit: transparency gap, greenwashing perception, profit motive, competitive strategy	Participant B: doubts how sustainable and transparent brands really are, and believes “any brand hides some facts because they cannot be perfect.” Participant E: describes fast fashion's constant new collections as creating a “greenwashing vibe.” Participant C: suspects sustainability claims are used for competitive edge. Participant F: views H&M's steps positively as an industry example.
Economic & Behavioral Barriers: self-reflection, incentive misalignment, price sensitivity	Participant B: wants to improve but sees room for growth. Participant A: values the recycling program and its discount, but names price as the dominant factor in purchase decisions.
Operational Friction & Awareness: limited knowledge, convenience vs. habit, blind faith	Participant F: appreciates garment-care guidance but has limited specific knowledge of H&M's initiatives. Participant C: donates clothing to charity rather than using H&M's in-store collection.

Table 3. Coded excerpts on consumer attitudes toward H&M's sustainability communication, drawn from focus groups conducted in Sweden (Zaidi)

Theme & Codes	**Representative Participant Excerpts**
Skepticism & Trust Deficit: distrust, validation seeking, marketing motivation	Fg4 P3(A): does not believe H&M is doing anything to protect the environment. Fg4 P3(C): sustainability claims make her more willing to buy from H&M. Fg4 P4(D): finds H&M's animal-welfare claims credible.
Economic & Behavioral Barriers: cost barrier, willingness to pay, personal priority	Fg4 P2(A): cost and limited variety make it hard to fully switch to sustainable products, though it remains a priority. Fg4 P3(D): willing to pay more for sustainable products.
Operational Friction & Awareness: material specifics, packaging concerns	Fg3 P1(B): describes H&M's claims about reusing material and organic sourcing. Fg4 P1(D) and Fg4 P3(B): raise concerns about plastic packaging undermining sustainability claims.

Discussion

Trust, Skepticism, and Their Cautious Implications for CBMI (Value Delivery)

Across the two secondary interview sources, we discern a common thread of skepticism surrounding H&M's ability to communicate sustainability effectively. Circular business models usually necessitate that customers be convinced that a certain action – such as returning a garment or selecting an item made of recycled material – results in an improvement in the environment. However, the documented skepticism may present an unseen challenge to consumers engaging with the circular services offered by H&M.

A participant from Ikonomou's research mentioned that she sensed a "greenwashing vibe" in messages offered by the fashion industry regarding sustainability. This statement shows the dissonance between a business model characterized by the high frequency of product introduction and the narrative around sustainability (Ikonomou). Taken at face value, the comment does not concern H&M's particular circular practices but only comments on sustainability in the fashion industry as such. The interpretation of the statement can be extended cautiously and point out that H&M emphasizes material longevity and recovery but adheres to the same business model that is characterized by the high frequency of product introduction and considerable sales volumes. Thus, if the skepticism observed in Ikonomou's study extends to H&M's circular communication, it speaks about a credibility issue faced by the company in the same vein as the one presented by the industry.

The findings from Zaidi's research respond to this approach. Different focus groups in Sweden had diverse opinions regarding H&M's claims regarding animal treatment and material compliance: some found the company's messages credible, while others rejected the idea of H&M having any commitment to environmental protection (Zaidi). This split, rather than uniform skepticism, is itself informative: it suggests that trust in H&M's sustainability claims is not settled among consumers, which would predict uneven rather than uniformly low participation in circular services if this general split extends to circularity specifically.

Economic Friction and Moral Licensing (Value Capture)

Both studies highlight the importance of price sensitivity as a roadblock to sustainable consumption. Price has been mentioned as a major factor when making purchasing decisions by two of Ikonomou's respondents. Some of Zaidi's focus group participants also referred to price as an important barrier in making the switch to sustainable products despite sustainability being a stated concern of many. If this phenomenon of price sensitivity applies to H&M's sustainable product line, it may indicate that a "green premium" that recycled-content clothing has may deter price-sensitive customers from making purchases of H&M products.

A more specific and more directly relevant pattern concerns H&M's recycling-for-discount program. One participant in Ikonomou's study described the discount she received for returning old clothing as “such a big thing,” explicitly framing it as an incentive to make a new purchase rather than as a replacement for one (Ikonomou). This pattern is consistent with what behavioral economics describes as moral licensing: performing a virtuous act, such as returning clothing for recycling, can create a sense of moral credit that makes a subsequent,

resource-intensive purchase feel more justified (Merritt et al.). Because H&M's own recycling program structurally pairs a purchase discount with the act of recycling, this is a case where the general behavioral-economics literature and a specific feature of H&M's business model align closely enough that the inference from general consumer behavior to CBMI-specific behavior is more defensible than in the trust-related findings above. If this pattern holds broadly among H&M's customers, its effect would be to increase, rather than decrease, garment turnover, working against the emissions benefit the recycling program is intended to produce.

Awareness and Operational Friction (Value Creation)

To make textile-to-textile recycling a viable option for H&M, a steady supply of used raw materials is necessary. The interview participants had opposing ideas about H&M's sustainability efforts being too vague to be trustworthy. The participant of Ikonomou's research explained only that H&M has “done something” in the textile recycling area and could not provide any additional details, while the participants of Zaidi's focus groups raised the issues of plastic packaging which are not directly linked to textile recycling yet show that that cognitive dissonance concerning the promotion of the brand at the packaging level could remain in customers’ minds and prevent them from trusting H&M regarding its wider sustainability offer. Due to the fact that giving clothes to charitable organizations is a well-known and socially accepted way of disposal, it can continue to compete with H&M's in-store textile collection points even among customers with a fair concern for the environment, which can be drawn from the general awareness results.

Absolute versus Relative Decoupling: The Rebound Effect and the Recycling Paradox

H&M's value-creation pillar shows the clearest technical progress. As Table 1 shows, the company reported 89% of materials from recycled or sustainably sourced inputs in 2024, and it has invested in Syre, a joint venture aimed at scaling textile-to-textile polyester recycling — a genuine attempt to move beyond open-loop recycling. Shifting from virgin, oil-based polyester to recycled fiber is a meaningful step toward reducing Scope 3 emissions.

There is evidence to suggest that this technological advancement has not managed to bring about a measurable and absolute decrease in environmental impact. This observation can be attributed to the rebound effect, which is derived from environmental economics. The rebound effect refers to a phenomenon whereby an increase in efficiency which is attributed to an innovation reduces the cost of consumption in either practical or psychological terms, leading to further consumption. In time, additional consumption could undo previously gained environmental benefits of technology (Sorrell). In terms of H&M and its recycling in exchange for a discount program mentioned during the interviews, it is easy to see how the rebound effect could work in this case as well as in relation to the overall relation between the efficiency of waste management processes and the total number of products sold.

The latter is aggravated by what some recent studies call the recycling of textiles paradox (Brydges). The majority of fast fashion recycled polyester comes from PET bottles instead of

clothing, which is an open loop operation that diverts waste while only postponing it for the duration of one product life cycle. Mechanical recycling is the dominant method of commercial use and does not provide quality material, as the very process of it reduces the length of polyester fibers and, as a result, requires the addition of virgin polyester fiber to be able to provide good quality fabric. Unless a method of chemical recycling is developed and commercialized – an aim that Syre pursues unsuccessfully on a large scale – the industry will keep recycling downcycling.

Governance, Scale, and Shared Responsibility

When considering the evidence presented above, it can be concluded that consumer behavior does not fully explain the environmental performance gap for H&M. H&M has its own production target, structure of recycling incentives and makes decisions on the pace and communication of its material transition. Thus, attributing circularity gaps to consumer behavior ignores the importance of these supply-side decisions.

The reverse issue is equally problematic: treating corporate decisions as the main cause of the gap. Fashion consumption takes place within social and cultural processes, such as the need for constant novelty, as well as the phenomenon of low-cost disposability of cheap clothing and consumption pressure, which have existed long before any retailer came up with its business model. Based on the findings of consumer motivation studies conducted through secondary interviews, it becomes clear that price, quality and social approval remain the key factors in making fashion purchases (Ikonomou). However, even if consumer behavior is partly dictated by decades of advertising in the fast-fashion segment, one can hardly distinguish the causes and effects of consumer choice and corporate strategy. The evidence reviewed here is more consistent with a shared-responsibility account, in which corporate production-scale decisions set the outer bound of what circularity gains are achievable while consumer norms and behavior determine how much of that bound is actually reached, than with an account that assigns responsibility entirely to one side or the other. Future research, discussed further below, would be needed to disentangle these contributions empirically.

Limitations

This research has several limitations that bound the claims made above. First, and most importantly, the two interview studies used here (Ikonomou; Zaidi) examined general consumer attitudes toward fast-fashion sustainability and H&M's sustainability communication, not H&M's CBMI specifically. This paper has attempted to flag this gap explicitly wherever interview evidence is extended to circularity-specific claims, but the underlying limitation cannot be fully resolved through careful framing alone; a study designed around CBMI-specific interview questions would be necessary to test the propositions above directly.

Secondly, both of the studies have utilized small scale of non-probability samples collected back in 2020 and 2021 in Scandinavia — six participants in Denmark and twenty participants in Sweden — predating H&M's more recent investments in Syre and expansion of Sellpy. The situation may have changed since the above research was conducted, as consumer

attitude towards the issue may have shifted and the data from samples are limited demographically and geographically considering H&M's huge and growing customer base, especially in Asia and America, where the dialogue about circular economy may be not as advanced as in Europe and North America.

Third, this paper does not conduct primary life-cycle assessment (LCA); it relies on H&M's self-reported KPIs alongside independent, published LCA and decoupling literature for benchmarking. Since this paper utilizes a secondary-analysis approach, the absence of product-line-specific LCA of H&M's garments makes comparisons between H&M's disclosures and external benchmarks remain indicative rather than conclusive.

Fourthly, this is a single case study. Even though the comparison with the competitors indicates the necessity for H&M to analyze its decoupling approach, there is no need for a multiple case study in this paper to prove the assertion that the company has the latest decoupling practices.

Future work based on this paper should focus on getting private information from H&M sustainability officers and on carrying out consumer surveys targeting CBMI from geographical perspective, independent LCS on certain H&M product lines and on completing a comparative case study exploring the similarities.

Conclusion

H&M Group's implementation of Circular Business Model Innovation is a significant strategic response to negative effects of fast fashion on the environment, and overall the findings presented indicate that CBMI serves effectively as both competitive strategy and regulation. It allows the firm to position itself in a competitive way just ahead of upcoming EU regulations in the sphere of apparel industry and produces real, quantifiable, and disclosed results in terms of recycled materials procurement and operational emissions reduction over five years. However, the findings do not confirm the statement that CBMI allows achieving the absolute reduction in environmental impact necessary for compliance with H&M's environmental aims in terms of Scope 3 emissions, which constitute the vast majority of the firm's carbon footprint. The findings show the opposite trend in spite of better performance. The gap in performance is due to various dimensions of corporate level decisions such as decisions regarding company's production volume, its incentives mechanism, and the pace of transition to closed-loop recycling, as well as consumer behavior trends including price sensitivity, trust issues, and low awareness supported in the literature.

Based on this evidence, three clear changes seem likely to reduce the gap between H&M's relative and absolute decoupling: (1) linking a larger portion of executive pay to material-recovery rate and absolute emissions instead of just revenue growth. This aligns with the governance issues highlighted by Guldmann and Huulgaard; (2) increasing service-based revenue streams, such as resale through Sellpy, garment repair, and rental, beyond their current pilot-level share of total revenue; and (3) combining the EU's Digital Product Passport

requirements with internal emissions targets tied to total production volume, focusing on absolute emissions rather than just per-unit emissions.

Overall, the findings suggest that Circular Business Model Innovation should be viewed as an important component of corporate sustainability rather than a standalone solution to environmental degradation. While H&M demonstrates meaningful progress toward relative environmental decoupling, achieving absolute sustainability will likely require complementary advances in production systems, governance, regulatory policy, and consumer behavior. The study therefore highlights the importance of evaluating corporate sustainability initiatives not only by their adoption but also by their measurable environmental outcomes.

Works Cited

Brydges, Taylor. "Closing the Loop on Take, Make, Waste: Investigating Circular Economy Practices in the Swedish Fashion Industry." Journal of Cleaner Production, vol. 293, no. 126245, Apr. 2021, https://www.sciencedirect.com/science/article/pii/S0959652621004650.

Conca, Lavinia, et al. "Exploring the Impact of Circular Economy Practices on Corporate Performance: A Global Analysis." Business Strategy and the Environment, Wiley, Nov. 2025, https://doi.org/10.1002/bse.70315.

European Commission. EU Strategy for Sustainable and Circular Textiles. 2022, https://environment.ec.europa.eu/document/download/74126c90-5cbf-46d0-ab6b-60878644b395_en?filename=COM_2022_141_1_EN_ACT_part1_v8.pdf.

European Environment Agency. "Greenhouse Gas Emissions from EU's Textiles Consumption." Europa.Eu, 27 Feb. 2025, https://www.eea.europa.eu/en/circularity/sectoral-modules/textiles/greenhouse-gas-emissions-from-eus-textiles-consumption.

Geissdoerfer, Martin, et al. "The Circular Economy - a New Sustainability Paradigm?" Journal of Cleaner Production, vol. 143, no. 1, Feb. 2017, pp. 757–68, https://doi.org/https://doi.org/10.1016/j.jclepro.2016.12.048.

Guldmann, Eva, and Rikke Dorothea Huulgaard. "Barriers to Circular Business Model Innovation: A Multiple-Case Study." Journal of Cleaner Production, vol. 243, no. 1, Jan. 2020, p. 118160, https://doi.org/10.1016/j.jclepro.2019.118160.

H & M Hennes & Mauritz AB. H&M Group Annual and Sustainability Report 2024. Mar. 2025, https://hmgroup.com/wp-content/uploads/2025/03/HM-Group-Annual-and-sustainability-report-2024.pdf.

Hong, Yuanxiao, et al. "Decoupling Carbon Emissions from Economic Growth in China's Textile Industry: A Decomposition of Driving Factors." Frontiers in Environmental Science, vol. 13, Frontiers Media SA, May 2025, https://doi.org/10.3389/fenvs.2025.1587498.

Ikonomou, Natalia. "Exploring the Gap Between Consumers' Sustainable Attitudes and Behaviors in Fast Fashion: A Case Study on H&M." CBS Research Portal, 2020, https://research.cbs.dk/en/studentProjects/exploring-the-gap-between-consumers-sustainable-attitudes-and-beh/.

Kirchherr, Julian, et al. "Conceptualizing the Circular Economy: An Analysis of 114 Definitions." Resources, Conservation and Recycling, vol. 127, no. 1, Dec. 2017, pp. 221–32, https://doi.org/https://doi.org/10.1016/j.resconrec.2017.09.005.

Merritt, Anna C., et al. "Moral Self-Licensing: When Being Good Frees Us to Be Bad." Social and Personality Psychology Compass, vol. 4, no. 5, May 2010, pp. 344–57, https://doi.org/10.1111/j.1751-9004.2010.00263.x.

Motke, Francies Diego, et al. “Circular Business Model Innovation and Its Relationship with Business Performance in Brazilian Industrial Chemical Companies.” Frontiers in Sustainability, vol. 2, Jan. 2022, https://doi.org/10.3389/frsus.2021.766696.

Niinimäki, Kirsi. “The Environmental Price of Fast Fashion.” ResearchGate, Apr. 2020, https://www.researchgate.net/publication/340635670_The_environmental_price_of_fast_fashion.

Ostermann, Cristina M., et al. “Business Model Innovation for Circular Economy in Fashion Industry: A Startups’ Perspective.” Frontiers in Sustainability, vol. 2, Oct. 2021, https://doi.org/10.3389/frsus.2021.766614.

Sorrell, S. The Rebound Effect: An Assessment of the Evidence for Economy-Wide Energy Savings from Improved Energy Efficiency. 1 Jan. 2007, https://www.researchgate.net/publication/258260565_The_Rebound_Effect_An_Assessment_of_the_Evidence_for_Economy-Wide_Energy_Savings_from_Improved_Energy_Efficiency.

Zaidi, Umar Usman. Analysis of Consumers” Attitudes Towards Sustainability Communication of H&M . 2021, https://www.diva-portal.org/smash/get/diva2:1594131/FULLTEXT01.pdf.

In What Ways Can Kaitiakitanga Principles Provide Guidelines for Renewable Energy Projects in Te Tai Tokerau? By Aston Plunkett

Abstract

For much of the twentieth century, an illegal pattern of land seizure and desecration exploited Māori resources at the hands of the Crown. Renewable Energy expansion in Aotearoa New Zealand was defined by a "legal blanket", systemically isolating Māori from their resources, only allowing *rangatiratanga* and *kaitiakitanga* where there are holes in the blanket to poke through. This research traces the legal and historical story of Māori energy sovereignty, investigating breaches of the Treaty of Waitangi, and identifying case studies that exemplify a productive partnership. Analysis is conducted through the lens of *kaitiakitanga,* a multi-dimensional, holistic Māori epistemology of environmental conservation, essential for Māori to exercise authority and protection over their resources. By synthesizing evidence from Waitangi Tribunal reports, legal analysis, Māori scholarship, and technical aspects of renewable energy, it provides a framework of *kaitiakitanga*-informed recommendations, a product of learnings from New Zealand's flawed past, and an opportunity to improve for a prosperous future. It discusses the specific situation of Te Tai Tokerau's impoverished, remote, rural communities with poor infrastructure - prime targets for renewable energy developments to spark a change in community wellbeing and uplift many out of poverty. This study centers Māori principles of *mauri, manaakitanga, tapu, noa, whakapapa,* and *mana,* functioning as guiding principles to confront the institutionally racist history of Māori exclusion. Ultimately, it aims to fill the gap of Māori, *kaitiakitanga*-informed guidance for renewable energy advancements in one of New Zealand's most deprived, rural, Māori regions.

Introduction

Aotearoa New Zealand is a global leader in renewable energy, with 85.5% of electricity and 45.5% of primary energy sources being provided from renewable sources (Ministry of Business, Innovation and Employment 1). The main methods of renewable generation in New Zealand consist of hydro, wind, geothermal, and solar. Expanding and further developing renewable energy in Aotearoa is crucial in order to meet its aspirational targets for greenhouse gas emissions.

New Zealand's large, tribally (*iwi*) based indigenous Māori population are significant stakeholders in renewables, commonly located on spiritually and culturally significant Māori land. Māori consider themselves to be the *kaitiaki* (guardians) of their specific ancestral lands, with a responsibility to preserve and protect them for future generations. The Māori concept at the core of this paper, *Kaitiakitanga,* is analysed as a responsibility to enhance *mauri,* uphold *mana* and *manaakitanga,* determined by strong *whakapapa* (genealogical connection). This transgenerational mandate ensures *mana*, the authority or control over the management of natural resources. Under this model, *mauri* draws links from biological and cultural outcomes to environmental management, representing an ecosystem's life-sustaining force, and

manaakitanga contributes reciprocal and unqualified acts of giving, caring, and hospitality (Harmsworth and Awatere 275-76). *Kaitiakitanga* includes a way of thinking and behaving to achieve spiritual, political, and physical well-being. (Marsden and Henare 3).

However, it is important to note that Western concepts fail to appropriately capture its multi-dimensional properties, commonly falling short and condensing its complexity to a single word: guardianship (Marsden and Henare 15). Interviewed by the NZ Treasury in 2020, Dr Shaun Awatere details that *kaitiakitanga:*

"deals with issues of connectedness between people and the environment and people to each other, and the responsibility that people have to look after the environment" (Awatere 1).

This paper aims to provide solid, practical guidance to upcoming renewables projects in Northland, as limited guidance is available to ensure the appropriate use of Māori lands and *taonga* (valuable treasures) for energy generation. Specifically, it identifies what forms of energy technology, governance structures, and project frameworks can best consistently uphold *kaitiakitanga.*

Te Tai Tokerau, home to a larger Māori population (37.4%), has much higher indicators of socioeconomic deprivation compared to the New Zealand average (Stats NZ, *Place and Ethnic Group Summaries*). Northlanders face lower household and average earnings paired with higher electricity prices, contributing to the nationwide issue of energy poverty, which is especially prevalent in remote, rural Northland communities (Infometrics). However, Northland holds geothermal resources and extensive solar potential (Stats NZ, *Place and Ethnic Group Summaries*). Renewable energy alternatives decrease power costs, reduce air pollution, and create more jobs when compared to fossil fuel generation (United Nations). The potential economic development renewables provide could be key in lifting Northland *whānau* (family) out of poverty and deprivation.

In Te Tai Tokerau, new renewable energy developments are on the rise, adding to an existing geothermal plant in Ngāwhā, which is the subject of controversy and breaches of the Treaty of Waitangi, to be discussed in detail later. The first grid-scale solar farm, Kohirā, finished construction in 2023, leading the way for a boom of solar developments recently completed or under current development. This paper is limited to solar and geothermal resources, as they are the only methods that have significantly established themselves in the region.

Given Northland's significant Māori population and connection to the land, it is more important than ever to ensure that renewable energy projects are completed with and under the guidance of *iwi* (tribes), for mutually beneficial outcomes.

This paper argues that Māori-led local energy operations, supported by culturally aligned renewable technologies and genuine authority over land and *taonga,* are critical to upholding *kaitiakitanga* and combating the historical inequities of New Zealand's past.

Methodology

Sources were selected by their relevance to *kaitiakitanga*, inclusion of a *Te Ao Māori* (the Māori world) worldview, and relevance to renewable energy in Aotearoa, ideally Te Tai Tokerau. Credible and recent sources were prioritised, as well as those written by Māori academics and trusts. Case studies were selected by their relevance to *kaitiakitanga* and renewable energy in Te Tai Tokerau. The quality of sources was assessed to ensure useful and reliable information was obtained, including: rigor of methodology, source credibility, Māori epistemologies, and peer-reviewed status.

This methodology aims to identify patterns across energy type, framework or governance system used, *iwi* involvement, and *kaitiakitanga* frameworks when applied. However, limitations resulted in the inclusion of news articles for information, as there is limited information available on current energy projects in Aotearoa New Zealand and their governance, history, and aims, especially when these projects involve *iwi*. Published interviews were publicly available and used to add a more direct Māori scholar input to the *kaitiakitanga* framework.

Literature Review

Key Findings

Legal and Historical Foundation

The repealed 1953 Geothermal Energy Act allowed the Crown to uplift Māori land to serve its own purposes in energy generation (Geothermal Energy Act). Conversely, Te Tiriti o Waitangi provides Māori with rights to exercise *kaitiakitanga*, but these have been subject to repeated and intentional breaches (Charters and Whare 13). While *kaitiakitanga* is cemented in the Resource Management Act (1991), it is often interpreted by the judiciary as an optional concern rather than a mandate (Keown 81). *Rangatiratanga*, supported by the legal principle of *contra proferentem*, remains the key for Māori to exercise authority over their resources (*Te Tiriti o Waitangi;* Keown 74-75).

Conceptual and Cultural Frameworks

Pākehā and Māori sources typically disagree on definitions of *kaitiakitanga*: the former sticking to a primitive definition of "guardianship," while Māori academics criticize this as minimizing the concept's multifaceted nature (Marsden and Henare 15). Tokenism is a recurring concern; the Crown and Pākehā scientists often deploy *kaitiakitanga* as a cultural "buzzword" without acknowledging its depth (Awatere 4). Māori share a deep spiritual connection with geothermal resources, considering them a sacred *taonga* to be protected (Waitangi Tribunal 1).

Energy Trends and Case Studies

Geothermal power dominates the literature for *iwi* engagement, as a large proportion of New Zealand's geothermal resources are located on Māori land (Tauhara North No.2 Trust, 1). Māori energy involvement was initially rare and unfair, but trends since 2010 show increasing

iwi involvement and productive partnerships. Māori energy ventures typically rely on corporate partnerships to provide capital and experience for large-scale projects (Tauhara North No.2 Trust 1).

Regional Solutions for Te Tai Tokerau

Energy poverty is a critical issue in Northland, exacerbated by poor infrastructure and lower socioeconomic indicators (Infometrics). Local renewable energy counteracts energy poverty and aligns with Local and Community Energy (LCE) objectives (Berka et al. 174). Micro-grid generation on *papakāinga* (communal Māori land) and *marae* (Māori meeting grounds) is ideal for exercising *rangatiratanga* and *kaitiakitanga* (Berka et al. 174). Agrivoltaics represents a promising approach to align solar generation with *kaitiakitanga* (Round, Rittick Maity et al. 1).

Findings

Governance

"These notions overlap with Māori conceptualisations of sustainable development that underpin iwi energy initiatives, underpinned by kaitiakitanga (collective guardianship of natural resources) and tino rangatiratanga (self-determination)." (Berka et al. 174)

Table 3
Characterisation of local and community energy profiles in New Zealand.

Profile	Legal status	Activities	Motivation	Level of engagement	Outcomes
Iwi trusts	Settlement trusts	Generation	Normative (Self-sufficiency, self-empowennent); Instrumental (Employment, community development)	Consent, Voting rights, Partnership	Education and employment for Iwi beneficiaries
	Charitable Trust	Energy efficiency, Generation	Instrumental (Health, employment, community development)	Iwi control, consultation	Local community improved services, education, employment for local Māori.

Figure 1: Characterisation of local and community energy profiles in New Zealand (Berka et al. 172)

Local and community energy (LCE), typically focusing on energy efficiency and power generation, is significant in the Māori energy economy, with 7.6% of LCE projects identifying as Māori-owned, operated by *iwi, hapū* (sub-tribe), *rūnanga* (tribal council), or Māori charitable trusts (Berka et al. 174). Māori-led initiatives often take a step further into energy efficiency programmes. Common motivations behind *iwi* LCE include addressing energy poverty among Māori communities, as well as generating education and employment opportunities for vulnerable members.

However, empirical studies show that LCE outcomes are not universal and depend on leadership motivation, degree of community engagement, business model design, and local context (Berka et al. 166).

Research identifies three main types of shared governance structures between communities and developers in renewable projects: Split ownership, Joint venture, and shared venture. In split ownership, the community and developer both provide different assets, with the

developer covering the majority of development costs. Joint ventures consist of each party sharing in electricity production, with the developer fronting costs of development whilst both parties own fixed proportions of assets. In a shared venture, the developer owns the assets and covers development costs, but shares revenue with the community (Tauhara North No.2 Trust, 4). Typically, the developer covers pre-construction costs, transferring risk away from communities and attracting risk-averse investors post-construction (IEA-RETD iv, 17).

As investors, Māori are typically cautious, opting to be risk-averse due to the intergenerational nature of their assets, preferring long-term stability. This is best supported by minimising capital investment while retaining control, with high Māori ownership being most feasible where technical and economic viability has already been demonstrated (Tauhara North No.2 Trust 4). Due to the complicated consenting and approval process, careful legal planning and shareholder consultation is required in the current governance landscape (Tauhara North No.2 Trust 4).

Legislation

"It wasn't long before the Crown (government) enacted colonial rule to bypass the genuine intent of Te Tiriti (The Treaty) to acquire Māori land by deception."(Tauhara North No.2 Trust 1)

Te Tiriti o Waitangi is the founding constitutional document of New Zealand. Written in both English and Te Reo Māori, significant differences exist between each version, signed by *rangatira* (Māori chiefs), and British representatives. The differences have created enduring challenges in governance, policy, and public tensions, compounded by deliberate Crown breaches of the agreement (Charters and Whare 13). Governance structures are guided by Te Tiriti, *rangatiratanga*, and long-term community wellbeing (Tauhara North No.2 Trust 2).

Derived from Te Tiriti o Waitangi are multiple principles, including mutual benefit, active protection, partnership, and *tino rangatira*, all relevant to renewable energy generation and subject to past breaches (Te Tiriti o Waitangi, 1840; Waitangi Tribunal). A source of much controversy throughout the treaty's history, *Rangatiratanga* is a crucial prerequisite for achieving *Kaitiakitanga,* as governance power and authority enables meaningful stewardship (Keown 75).

Widely accepted into legal and academic spaces is the principle of contra proferentem ("interpretation against the draughtsman"), which prioritises Te Tiriti o Waitangi (Te Reo Māori version) over the English text to resolve ambiguity (Suter 6-7, 18).

Established in October 1975, the Waitangi Tribunal exists to address Crown breaches of Te Tiriti o Waitangi, awarding settlements, including formal apologies and assets to aggrieved parties (*Treaty of Waitangi* Act 1975).

Geothermal

"Māori are key participants in Aotearoa (New Zealand) geothermal projects through their ownership of land that provide access to world class geothermal resources." (Tauhara North No.2 Trust 1)

Geothermal springs and resources are of significant interest to the Māori economy, with an increasingly large proportion of geothermal generation being located on Māori land (Tauhara North No.2 Trust 1). Māori-owned geothermal enterprises vary in their governance structure, varying from 100% *iwi* ownership to joint ownership, but typically take the shape of the latter (Berka et al. 174). Joint ventures with *iwi* are common practice as well as community engagement, voting, and local *iwi* advising (Berka et al. 174; Wilson et al. 2). Māori-owned geothermal enterprises commonly leverage reparations from Treaty settlements to acquire geothermal businesses (Berka et al. 174).

Located in Rotorua, the Tauhara North No.2 Māori Land Trust (TN2T) partially owns three geothermal power plants in partnership with Mercury, a majority state-owned power company. A productive partnership between the two exists, with technical expertise and capital provided by Mercury, whilst TN2T provides their land and support with consenting processes (Nga Tamariki Geothermal Station).

Nearby TN2T's expansions, the Tuaropaki Trust holds 75% shares of their Tuaropaki Power Company, selling their energy under a wholly debt financed geothermal generation model in partnership with Mercury (Tauhara North No.2 Trust 3; Simpson 17).

The Ngāwhā geothermal field, of particular interest to this study, is a geothermal power station and geothermal pools. The *iwi* trust has lost autonomy over the geothermal power generation to Top Energy, but remains the owner of one singular acre from an originally 5097 acre block, and has oversight over the geothermal pools (NRC).

An innovative example of *iwi* energy governance exists in the Bay of Plenty, where Ngāti Tūwharetoa Settlement Trust (NTST) are the only wholly *iwi*-owned geothermal supplier in New Zealand (Berka et al. 174.) In alignment with *Kaitiakitanga,* revenue from the geothermal operation is recycled to the local community, providing scholarship, elder support, dividend returns and beneficiaries, explicitly aiming to build "*identity and pride in the land and potentially some jobs in the area"*, while avoiding *"stepping into the breach of the Crown"* (Berka et al. 174).

Solar

"Agrivoltaics is said to have many potential benefits, including less stress on the livestock and crops, with the panels providing shade and protection, improved productivity and added revenue for the farmer amid the ups and downs of weather and commodity prices (Round)."

Solar generation has the advantage of being scalable, adapting to smaller operations on houses or marae, or large grid-output operations. Data shows a strongly increasing trend in solar generation in Te Tai Tokerau, with three large-scale solar farms being constructed in 2023 and 2024, compared to none prior to 2023. These projects have all been undertaken in partnership or with oversight from local *iwi.*

Lodestone Energy collaborated with local *iwi* on New Zealand's largest solar farm, Kohirā. Gifted by Ngāti Rarawa, indicating a respectful, culturally appropriate partnership, Kohirā translates to Suncatcher. *Iwi* were consulted in the process, creating a mutually beneficial

project, jobs for locals in Kaitaia (many of whom are *iwi* members), and restoring 7 hectares of native forest. (Kohirā Solar Farm)

Kohirā is also an innovative example of agrivoltaics, with panels providing shade and protection for a flock of sheep. *"They're sort of content, they're full,"* claimed a local farmer on the Kohirā solar farm (Round).

Similarly, the Naumai solar project experiences strong *iwi* collaboration, from Ngāti Whātua Hapū Te Uri o Te Hau and Naumai Whanui. *"Collaborating with local iwi and engaging local contractors are the best ways to ensure the community continues to benefit long-term,"* stated a representative of NewPower, who partnered with Māori for the project. Neighbouring the Naumai solar project in Ruawai, Te Uri o te Hau *hapū* have partnered to produce another solar farm, this time with Northpower, a community-owned electricity supplier. *Iwi* gifted the name, Te Puna Mauri ō Omaru, of the solar farm, in a show of respect and good faith, indicating a productive partnership.

Research has shown solar generation can have a negative environmental impact: bird collisions and burns from solar rays from heat energy infrastructure suggest that avian impacts may be damaging (Upton). Notably, it has been proven that correct environmental planning significantly reduces the risk of these avian impacts.

Solar energy provides specific uses to *papakāinga*. *Papakāinga* describes housing on ancestral homes of Māori, being multiple-owned property that typically hosts long-term residents, making it ideally suited to the 5-7 year payback period that offputs many landlords and other property owners (Ison 70-71).

Mātauranga Māori and Kaitiakitanga

"Kaitiakitanga continues to find centrality in Māori kin-based communities because it weaves together ancestral, environmental and social threads of identity, purpose and practice."(Kawharu 349-350)

Mātauranga Māori is defined as "*the body of knowledge originating from Māori ancestors, including the Māori world view and perspectives, Māori creativity and cultural practices,"* by the Māori Dictionary ("Mātauranga Māori"). As it is a Māori concept, this paper does its best to capture the broad, multi-dimensional nature of its true meaning.

The depth of *Kaitiakitanga* as a concept has been explained by Māori academic Shaun Awatere:

"*There is a very strong core ethic there that natural resource management is about ensuring the sustainability and the livelihoods of your children, of your mokopuna and their children. It's probably the antithesis to the individualistic profit maximisation approach. The other concept connected to kaitiakitanga is whakapapa. The ability to carry out kaitiakitanga is very much dependent on whether you have that genealogical connection to an area or you've been living there for a certain amount of time." (Awatere 1)*

Māori believe that everything originated from *Ranginui* (Sky father)and *Papatūānuku* 's (Earth mother) union, with *kaitiakitanga* grounding itself in this relational perspective where people are privileged to use and exist alongside the land and its resources, as opposed to western models of ownership (Marsden and Henare 2,12,15,17,18). As such, only *tangata whenua* (indigenous people) may define the role of *kaitiaki* (Wilson et al. 1).

"From a Māori perspective, the origin of the universe and the world can be traced through a series of ordered genealogical webs that go back hundreds of generations to the beginning. This genealogical sequence, referred to as whakapapa, places Māori in an environmental context with all other flora and fauna and natural resources as part of a hierarchical genetic assemblage with identifiable and established bonds."(Harmsworth and Awatere 274).

Roles of *kaitiakitanga* include utilising *rahui* (restrictions) to regulate behaviour and protect *taonga*, and upholding Māori spiritual, social, and environmental obligations, thus appealing to tikanga. *Kaitiakitanga* is intrinsically linked to *rangatiratanga, mana, mauri, tapu, noa,* and *manaakitanga.* (Harmsworth and Awatere 275-276). *Tapu*, *"Sacred or restricted access to resources, or places, for customary and cultural reasons"* (Harmsworth and Awatere 285), creates restrictions to manage and protect the relationship between humans and the environment. Similarly, *noa* guides interactions with our sacred resources, existing as *"Open access to resources – but under kaitiakitanga practice (still protected, valued, managed)*" (Harmsworth and Awatere 284).

In alignment with *kaitiakitanga*, it is important to ensure duty of care to the environment (*Te Taiao*), land (*whenua)*, and people (*tāngata*), which guides decision-making in development projects (Tauhara North No.2 Trust 1).

1.	Whanaungatanga: relationship-building with iwi.
2.	Opportunities for iwi participation in planning and development.
3.	Awareness of *tangata whenua* values in project teams.
4.	Iwi continuously informed on project developments.
5.	Agreed methodology for engagement at appropriate levels.
6.	Iwi engagement with consistent personnel.
7.	Face-to-face engagement (Kanohi ki te kanohi).
8.	Visible participation alongside iwi (Kanohi kitea).
9.	Adherence to kawa (protocols), tikanga (customs), kaupapa (underlying philosophy).
10.	Focus on kōrero (spoken word).
11.	Relationship Agreement: recognition of Treaty, statutory obligations, principles, consultation processes, conflict resolution, and cultural impact assessments.

Table 1: GNS Kaitiakitanga Recommendations (Wilson et al. 2).

Kaitiakitanga is now becoming increasingly embedded in national funding schemes and scientific discourse (McAllister et al. 2). The current research landscape has an increasing group of successful and respected co-developed projects integrating both *mātauranga* and science.

Additionally, the importance of *kaitiakitanga* in Māori environmental governance is displayed by its frequent recurrence in *iwi* environmental plans (McAllister et al. 2).

Historic Māori Exclusion

"*The Act simply ignores any such Treaty rights and gives the Crown the legal right if it so chooses, to act without regard to such rights."(Waitangi Tribunal 141)*

The 1991 Resource Management Act recognises Māori's relationship, culture, and traditions to *" ancestral lands, water, sites, waahi tapu, and other taonga:", " particular regard to kaitiakitanga",* and *"the principles of the Treaty of Waitangi"* (*Resource Management Act* ss. 2, 7). The *"particular regard"* to *kaitiakitanga* is interpreted as intentionally optional rather than a mandatory obligation to uphold it (Keown 80; *Resource Management Act* ss. 2, 7). The overall focus of the RMA dwarfs its mentions of *kaitiakitanga* and Māori as recognised *kaitiaki* of the land, highlighting the need for larger government and legislative recognition, enabling productive, culturally appropriate projects to be protected and respected under New Zealand law.

The now-abrogated 1953 Geothermal Energy Act legalised the *"Sole right to tap and use geothermal energy to vest in the Crown"* (Geothermal Energy Act). The act blatantly removes the right of self-determination for Māori over their own land, systemically dispossessing them of a potential income source and their right to exercise *kaitiakitanga*. This highlights the contrasting perspectives: the Crown viewing geothermal resources as a financial opportunity, whilst Māori perceive them as *taonga* to be preserved and respected.

Case studies from the Central North island highlight this historical disconnect. The 1950s Wairakei geothermal project was void of meaningful Māori input, despite a strong cultural and spiritual significance of the area to local Māori. Even in the 1990s, the Ohaaki project was yet again an example of Māori underrepresentation, with a peppercorn lease of $1 annually being offered to Ngāti Tahu *iwi* and restricted land access (Tauhara North No.2 Trust 3).

A blatant example of this, found in Northland, is the 1993 Ngāwhā Geothermal Resource report, highlighting the cultural significance of the resource to Ngāpuhi *iwi*. The report describes: *"In traditional Māori terms, the springs are the face or eye of the resource, but its whatumanawa or heart is below the ground and connected to other surface manifestations in the district"* (Waitangi Tribunal 1).

This report establishes the depth of connection *iwi* have to the Ngāwhā resource - and in doing so conveying the profound hurt experienced by culturally inappropriate and improper extraction of the geothermal resource for the generation of capital. A cultural impact report on the Ngawha geothermal operation supports this, stating *"by definition any unauthorised use of the energy and power of the Ngawha reservoir tramples the mana of the ancestors, the people of today and of those generations to come"* (Parahirahi C1 Trust 35). The Parahirahi C1 Trust and its affiliated *iwi*, although initially supportive of the geothermal developments, are affronted by the use of their *taonga* by Top Energy to generate electricity, risk the vitality and *mauri* of their reservoir, and share none of the economic returns. (Parahirahi C1 Trust 35)

Furthermore, the tribunal report found that this was in direct breach of both the Treaty of Waitangi and the Resource Management Act, clearly violating the rights of *iwi* to further their agendas. The report highlights scathing breaches of New Zealand law: "*the claimants' taonga may be placed in jeopardy entirely against their will"*(Waitangi Tribunal 153). In alignment with the principles of Te Tiriti and the Resource Management Act, the Waitangi Tribunal recommended that all 4 acres of the Ngāwhā geothermal zone be returned to Māori (Waitangi Tribunal 154). However, over 30 years later, it is council property and *iwi* only have autonomy over the thermal pools (Ngāwhā Generation).

Currently, many policy, legislative, and financial barriers still oppress Māori renewable developments. As a result of this, most standalone generation projects by local energy co-operatives (often Māori community energy projects) remain inactive or 'in limbo' (Berka et al. 175). LCE actors, seeking ways around barriers, are often forced to partner with incumbent utilities, resulting in unideal negotiations (Berka et al. 175).

LCE and grassroots energy actors have little evidence to support a legitimate reputation in the energy industry and are especially dependent on a favorable policy environment, making them vulnerable to current uncoordinated council processes and a lack of recognition in New Zealand policy (Berka et al. 179).

Misuses and Tensions of Kaitiakitanga

"The Pakeha system of ownership can be seen as a blanket with the Māori concepts of kaitiakitanga and rangatiratanga only operating in those areas where there are holes or where the blanket simply provides no cover." (Keown 65)

Table 2. Some differences between the concepts underpinning kaitiakitanga and conservation management but see Mace (2014) and Evans (2021) for a review of the ways conservation biology is evolving. Similarly, Bataille et al. (2021) provide specific examples of how landowners and Māori environmental guardians prioritise different values in wetland ecosystem management.

Kaitiakitanga	Western conservation
Not based on intrinsic value; traditionally managing for abundance (i.e. kai)	Based on intrinsic value of each component of the ecosystem
Based on a deep connection with the environment derived by whakapapa	Often driven by concepts of risk management or problem solving to avoid or fix environmental issues that are often caused by humans (such as invasive species)
Explicitly active management, including management of relationship with the environment	Includes a passive management component but can be active (e.g. control of invasive species, relocation of threatened birds to offshore islands, breeding intervention programmes)
Based on Mātauranga Māori	Based on Western science, highly data-driven
Holistic management of a system (ki uta ki tai, from the mountains to the sea)	Compartmentation/ fragmentation of systems, often with a focus on single species, although there has recently been a resurgence of Western conservationists recognising the value of a holistic view.
In te ao Māori, humans are part of the environment and Māori descend from it	Anthropocentric/separatist/preservationist
Has spiritual, inanimate, animate elements	Not spiritual
Entrenched in many core Māori values (i.e. whakapapa, reciprocity, rangatiratanga, mauri etc.)	Can be in conflict with Māori values (i.e. through preventing traditional harvest)
Underpinned by whakapapa and tikanga Māori	Governed by legislation and ownership

Figure 2: Differences between Western Conservation and Kaitiakitanga (McAllister et al. 8).

The legal "blanket" is indicative of the continued exclusion and degradation of Māori values in New Zealand legislation, highlighting the harsh contrast between Western and Māori approaches to ownership. Similarly, Western and Māori approaches to conservation, as discussed

above, contrast the active, holistic, relationship-based philosophy followed by *kaitiakitanga,* with the outcome-based, scientific, and more fragmented Western philosophy. Pākehā views are individualistic and heavily influenced by Eurocentric, Christian property norms, compared to interdependent, relational, and spiritually integrated Māori outlooks (Keown 68). Consequently, Western law prioritises individual ownership, diverging from Māori emphasis on interdependent stewardship of land and resources (Keown 68). Controversy has also emerged from the Te Tiriti principle of partnership providing political dialogue opportunities rather than measured action, commonly tokenising *kaitiakitanga* and *rangatiratanga* in the process (McAllister et al 2). Māori academics strongly caution against the "we're all *kaitiaki"* sentiment, an increasing trend, risking cultural appropriation and undermining mana whenua authority in projects, policy, and the wider New Zealand context (Ataria et al.)

Opportunities

"They [Maori-owned LCE] are also embedded in longer term community development strategies aimed at generating local socio-economic opportunities, reversing historical economic decline and restoring Māori ways of life." (Berka et al. 174)

Despite challenges, LCE and Māori energy initiatives have shown the ability to overcome institutional and financial barriers, offering many benefits to local communities. Social cohesion, local empowerment, and financial returns are all benefits of a community ownership model. Local businesses and trades contracted for renewable developments also contribute to strong networks, sparking local economic development, community pride, and transferable skills (Hicks and Ison 247).

In New Zealand, rural communities are particularly well suited to communal energy models, with shared vision, renewable energy access, and electricity affordability challenges making communities especially eligible (Armstrong 14). Northland especially aligns with this, home to a very high rural population of 48%, compared to a national average of 15% (Northland Regional Council). Research shows that bottom-up, community-led initiatives relying on engagement, education, and participation are critical to overcoming barriers (Berka et al. 167,174, 176).

In the LCE study by Berka, all three identified off-grid Māori projects demonstrated high degrees of community involvement in project design and implementation, political motivation for self-sufficiency, explicit aims of socio-economic development on ancestrally owned land, and commitment to sustainable use of Māori natural resources (Berka et al. 174). Additionally, Berka found that integrated microgrids remain at the feasibility stage, and are grant-funded in partnership with long-term community development strategies to generate local employment, reverse historical economic decline, and restore Māori ways of life (Berka et al. 174). Integrated microgrids could be especially useful for smaller-scale projects on *marae, papakāinga,* or remote rural locations, where Māori populations are significant and usually in situations with higher deprivation indicators.

Te Tai Tokerau Context

"Current regulations place the burden of infrastructure upgrades on the Northland community. However, the main beneficiaries of infrastructure development are likely to be the generation developers and the broader population, including Auckland." (MinterEllisonRuddWatts)

Te Tai Tokerau is subject to 13% higher average electricity costs than the rest of New Zealand (Northland Regional Council). The unemployment rate is 6.1%, compared to the national average of 5.0% (Infometrics). NEET, a measure of people aged 15-24 not in employment, education, or training, sits at 19.1% of Northland's population, compared to a national average of 12.9% (Infometrics). As of the 2023 census, internet access sits at 70.7%, compared to the Aotearoa average of 81.8% (Infometrics). These indicators clearly present Northland's economic and social situation as more deprived and less resourced compared to the rest of New Zealand.

Figure 3: Transmission and Distribution Network Ownership in Northland (Transpower, Northpower, and Top Energy)

Figure 4: Map of Northland Region's overall Deprivation (Yong et al.).

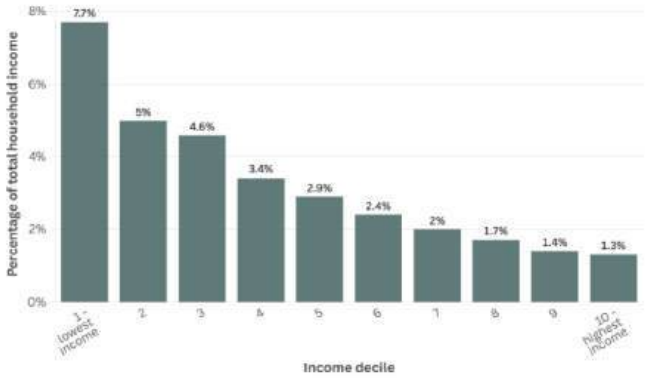


Figure 5: Electricity expenditure as a percentage of total income (O'Sullivan et al.).

As shown by the above diagrams, deprivation is highest in the regions with the lowest access to the electricity distribution network, highlighting the need for better distribution and access to electricity to combat energy poverty Northland also has a very large proportion of the Māori population, 37.4% compared to a 2023 census national average of 17.8% (Stats NZ). Energy poverty is a key issue, especially for remote Māori communities, which largely goes undocumented and unmeasured, but is acknowledged in council documents and local news articles regularly as an issue in dire need of fixing (Northland Mayors and Chairs Forum). Nationally, however, energy poverty is recognised and measured, with the poorest households spending the highest proportion of their income on electricity, as shown in the graph above. In terms of potential, Northland has very significant wind, solar, and geothermal resources that can be harnessed. However, this is very limited by the poor infrastructure, making it an unappealing location for projects, with high connection costs proving a barrier to innovation (MinterEllisonRuddWatts).

Despite this, the Ruakākā solar farm is in its development stage, planned for completion by late 2026, when it will produce enough to power around half the homes in Northland (Northland Regional Council). This makes the potential for economic development from renewable energy even more important for Northland as a region. This is a very significant development, aiming to increase supply and decrease electricity prices for Northlanders, and increasing the relevance of my research to the region. Ensuring *kaitiakitanga* alignment in a development as large as this is crucial for Northland, given its large proportion of vulnerable and Māori communities. Applying *kaitiakitanga* consistently will work to ensure cultural alignment and sensitivity to *Te Ao Māori,* whilst generating clean energy, economic development, and ultimately improving the wellbeing of Te Tai Tokerau's population.

Kaitiakitanga Recommendations

I have included recommendations for upholding each individual characteristic of *kaitiakitanga*, adding to the framework for Geothermal generation developed by GNS scientists. Each recommendation identifies a key feature of projects that are highly aligned with *kaitiakitanga*. Projects should aim to incorporate as many as feasible to illustrate a holistic assurance of *kaitiakitanga* throughout planning, construction, and production.

Whakapapa

Ensure local *iwi* input and oversight as only *tangata whenua* can be kaitiaki, including the monitoring of quality and use of resources, prioritising sustainable practices.

In terms of planning and development, focus planning on innovative, ambitious visions for the future, and generate solar energy on marae, papakainga, and other culturally significant locations to provide energy security for *rangatahi* (youth) and future generations.

Honor *whakapapa* to the land, its energy potential is considered local *taonga* rather than a commercial asset.

Rangatiratanga

Ensure permanent ownership for *iwi* over their land, including control over how their natural resources, energy extracted, and revenue generated are used. This includes phasing out reliance on developers and external providers of funding. Prioritise entire community engagement and collective decision-making, and ensure *iwi* control over how generated energy is distributed: *marae, papakainga, whānau* suffering from energy hardship, etc. Pursue smaller-scale projects that allow *iwi* full control in planning, construction, and distribution of energy to their local needs.

For current projects, prioritise split ownership and joint ventures over shared ventures, ensuring enhanced *iwi* control. Achieve fair, productive partnerships with culturally respectful and aware corporations providing expertise and capital where necessary, exemplifying positive bicultural approaches.

In terms of governance, acknowledge the Treaty principle of Māori self-determination as paramount in project *kaupapa* (plans). Pursue legal pathways through Te Tiriti o Waitangi and the RMA continuously to advocate for increased *rangatiratanga.*

Mana

Source local knowledge, trades, and expertise, contributing to community pride, *whanaungatanga*, and skill development, whilst giving priority to supporting communities' most vulnerable members - enhancing *mana* of the whole by tackling energy poverty.

Acknowledge *kaitiakitanga* holistically with respect to *Te Ao Māori*, and advocate for recognition and public awareness of historical wrongdoing by the crown - rightfully restoring *mana.*

Mauri

Maintain, respect, and honor spiritual ties to *taonga*, continuing to rehabilitate resources, land, and surrounding areas, reviving the life force.

In the planning phase, investigate and pursue agrovoltaics and other innovative pathways where feasible, and complete rigorous environmental planning to ensure avian impacts are limited.

In terms of future infrastructure, Prioritise developments to transition Northland into a resilient, resourced energy network, and pursue resilient, micro-grid projects for rural or remote projects.

Tapu and Noa

Avoid and leave culturally significant areas, specifically geothermal springs and fields of great cultural significance, out of respect. Respect *kaitiaki,* who are given full authority to manage and protect resources to enhance *mauri.*

Manaakitanga

Involve and listen to local stakeholders, whilst recycling financial returns into *iwi*-controlled initiatives: benefits, scholarships, further development, and investments etc.

Implement efficient, innovative energy to reduce waste and address energy poverty.

In practice, these principles deliver a culturally integrated, local community approach designed to spark economic development, increase socioeconomic outcomes for vulnerable, deprived communities, and inspire future developments to follow the same path of innovative cultural awareness. The environment is protected and respected, upholding not only *kaitiakitanga* but also *Te Ao Māori* and its holistic worldview.

Conclusion

New Zealand has a shameful history of institutionally depriving Māori of lands, resources, and spiritual *taonga* in the name of energy. Despite recent improvements, many errors of the past remain hurtful scars *iwi* still face today, damaging their trust and relationship with New Zealand government entities as remnants of the crown, and hampering future collaboration. In today's energy development landscape, technical support, capital, and protection from external entities remain essential for *iwi* to gain a seat at the table, highlighting the power imbalance and enduring barriers to *rangatiratanga.*

Combining *kaitiakitanga* with the technical, scientific depth of western conservation is an exciting prospect that has yet to be significantly explored in Te Tai Tokerau's energy sector, despite its obvious and unlimited potential for growth in the region.

The ambitious framework outlined by this paper elevates *kaitiakitanga* to a value inseparable from the energy sector's future. It envisions a Te Tai Tokerau bolstered by renewable energy development; a greener future where bicultural partnership makes way for *iwi* self-determination, producing vast quantities of electricity, distributing it equitably, and drastically improving lives through holistic support designed to lift *whānau* out of poverty and into a wealth of opportunities.

Works Cited

Armstrong, Alistair S. Sustainable Energy Management for a Small Rural Subdivision in New Zealand. 2009. Massey University, Master's thesis. Massey University Research Online, mro.massey.ac.nz/bitstreams/a153ccad-03d1-4dc9-b87a-564b836bbc8b/download.

Ataria, James, et al. "Whakamanahia te Mātauranga o te Māori: Empowering Māori Knowledge to Support Aotearoa's Aquatic Biological Heritage." New Zealand Journal of Marine and Freshwater Research, vol. 52, no. 4, 2018, pp. 467–86, doi:10.1080/00288330.2018.1517097.

Awatere, Shaun. "A Sense of Kaitiakitanga: Connectedness, Responsibility, People and the Environment." Interview by New Zealand Treasury. Feb. 2020, www.treasury.govt.nz/sites/default/files/2021-07/Treasury_LTFS%202020_Interview_A4_Interview%201_Dr%20Shaun%20Awatere_v2.pdf.

Berka, Anna L., and Emma Creamer. "Taking Stock of the Local Impacts of Community-Owned Renewable Energy: A Review and Research Agenda." Renewable and Sustainable Energy Reviews, vol. 82, 2018, pp. 3400–19, doi:10.1016/j.rser.2017.10.050.

Berka, Anna L., Janette L. MacArthur, and Claudio Gonnelli. "Explaining Inclusivity in Energy Transitions: Local and Community Energy in Aotearoa New Zealand." Environmental Innovation and Societal Transitions, vol. 34, 2020, pp. 165–82, doi:10.1016/j.eist.2020.01.006.

Charters, Claire, and Tamaki Whare. "Shaky Foundations." World Policy Journal, vol. 34, no. 4, 2017, pp. 11–14, doi:10.1215/07402775-4373434.

Geothermal Energy Act 1953. New Zealand Legislation, www.legislation.govt.nz/act/public/1953/0102/latest/DLM283204.html.

Harmsworth, Garth R., and Shaun Awatere. "Indigenous Māori Knowledge and Perspectives of Ecosystems." Ecosystem Services in New Zealand: Conditions and Trends, edited by J. R. Dymond, Manaaki Whenua Press, 2013, pp. 274–86.

Hicks, Alison, and Nicole Ison. "Community-Owned Renewable Energy (CRE): Opportunities for Rural Australia." Rural Society, vol. 20, no. 3, 2011, pp. 244–55, doi:10.5172/rsj.20.3.244.

IEA-RETD. Cost and Financing Aspects of Community Renewable Energy Projects: Volume II: German Case Study. Ricardo Energy & Environment & Ecologic Institute, 2016.

Ison, Nicole B. Opportunities and Barriers to, and Benefits and Impacts from, Papakāinga Owned Energy Systems: A Case Study of Parihaka. 2015. Massey University, Master's thesis. Massey University Research Online.

Kawharu, Merata. "Kaitiakitanga: A Maori Anthropological Perspective of the Maori Socio-Environmental Ethic of Resource Management." Journal of the Polynesian Society, vol. 109, no. 4, 2000, pp. 349–370. ResearchGate, www.researchgate.net/publication/284045145_Kaitiakitanga_A_Maori_anthropological_perspective_of_the_Mori_socio-environmental_ethic_of_resource_management.

Keown, Boyd. "Ownership, Kaitiakitanga, and Rangatiratanga in Aotearoa/New Zealand." Te Tai Haruru: Journal of Māori Legal Writing, vol. 2, 2006, pp. 64–83.

"Kohirā Solar Farm." Lodestone Energy, 6 Oct. 2024, lodestoneenergy.co.nz/kaitaia/.

Maity, Rittick, et al. "Agrivoltaic Systems for Sustainability: An Overview of Emerging Trends and Practices." Solar Compass, vol. 16, 2025, p. 100148, doi:10.1016/j.solcom.2025.100148.

Marsden, Māori, and Te Aroha Henare. Kaitiakitanga: A Definitive Introduction to the Holistic World View of the Māori. Ministry for the Environment, 1992.

"Māturanga Māori." Te Aka Māori Dictionary, www.maoridictionary.co.nz/word/12997. Accessed 19 July 2026.

McAllister, Tara, Daniel Hikuroa, and Cate Macinnis-Ng. "Connecting Science to Indigenous Knowledge: Kaitiakitanga, Conservation, and Resource Management." New Zealand Journal of Ecology, vol. 47, no. 1, 2023, article 3521.

Ministry of Business, Innovation and Employment. Energy in New Zealand 2025. Ministry of Business, Innovation and Employment, 2025.

MinterEllisonRuddWatts. "Renewable Energy Zones: Bridging the Gap for Northland's Electricity Resilience and Reliability." 25 Feb. 2025, www.minterellison.co.nz/insights/renewable-energy-zones-bridging-the-gap-for-northland-s-electricity-resilience-and-reliability.

"Nga Tamariki Geothermal Station." Mercury, 2026, www.mercury.co.nz/about-us/renewable-energy/geothermal/nga-tamariki-geothermal-station.

Ngāwhā Generation. "Our Stories." Ngāwhā Generation, www.ngawha.nz/our-stories.html. Accessed 19 July 2026.

Northland Mayors and Chairs Forum. Northland Regional Deal Proposal: A Shared Vision for Te Tai Tokerau. 2024, www.northlandnz.com/assets/Resources/docs/Northland-Regional-Deal-Proposal-Redacted-Full-Document.pdf.

Northland Regional Council. Chapter Ten – Northland Regional Council Annual Environmental Monitoring Report 2001-2002. Northland Regional Council, 2002.

---. "Economic Quarterly | December 2025." Northland Regional Council, 8 Dec. 2025, www.nrc.govt.nz/your-council/economic-development/economic-quarterly/2025/december/economic-quarterly-december-2025/.

---. "Economic Quarterly | October 2025." Northland Regional Council, 8 Oct. 2025, www.nrc.govt.nz/your-council/economic-development/economic-quarterly/2025/october/economic-quarterly-october-2025/.

O'Sullivan, Kimberley, et al. "Energy Poverty: The Lowest-Income Households Pay More in Aotearoa." Public Health Communication Centre Aotearoa, 13 Nov. 2024, www.phcc.org.nz/briefing/energy-poverty-lowest-income-households-pay-more-aotearoa.

Parahirahi C1 Trust. Cultural Impact Assessment: Expansion of the Ngāwhā Geothermal Power Station. 2014, http://www.ngawha.nz/uploads/3/2/1/2/32123857/final_011014_cultural_impact_assessment_parahirahi_c1_trust_1.pdf.

"Place and Ethnic Group Summaries: Northland Region." Stats NZ, 2024, tools.summaries.stats.govt.nz/places/RC/northland-region#new-zealand-index-of-socioeconomic-deprivation.

"Regional Economic Profile: Northland Region Overview." Infometrics, 2026, regions.infometrics.co.nz/northland-region.

Resource Management Act 1991. New Zealand Legislation, ss. 2, 7, www.legislation.govt.nz/act/public/1991/0069/388.0/DLM230271.html.

Round, Simon. "Kohirā Solar Farm: Northland Sheep Benefiting from the Power of the Sun." RNZ: Country Life, 5 July 2024, www.rnz.co.nz/national/programmes/countrylife/audio/2018945590/kohira-solar-farm-northland-sheep-benefiting-from-the-power-of-the-sun.

Stats NZ. "Ethnicity and Number of Ethnic Groups Specified by Age Group and Sex, for the Census Usually Resident Population Count, 2013, 2018, and 2023 Censuses." Stats NZ Tatauranga Aotearoa, 2023, explore.data.stats.govt.nz/vis?df[id]=CEN23_ECI_019&df[ag]=STATSNZ.

Suter, Blake. The Contra Proferentem Rule in the Reports of the Waitangi Tribunal. 2014. Victoria University of Wellington, Master's thesis. OpenAccess Te Herenga Waka Victoria University of Wellington, openaccess.wgtn.ac.nz/articles/thesis/The_contra_proferentem_rule_in_the_reports_of_the-Waitangi_Tribunal/17009864.

Tauhara North No.2 Trust. The Nga Awa Purua Geothermal Project, Rotokawa, New Zealand. 2024, media.business-humanrights.org/media/documents/Tauhara_North_No2_Trust_paper.pdf.

Te Tiriti o Waitangi [The Treaty of Waitangi]. 6 Feb. 1840, www.archives.govt.nz/discover-our-stories/the-treaty-of-waitangi.

Transpower, Northpower, and Top Energy. Resilience, Reliability and an Energy Bridge: Te Tai Tokerau Northland. 2024, static.transpower.co.nz/public/uncontrolled_docs/Resilience%20Reliability%20and%20an%20Energy%20Bridge%20Te%20Tai%20Tokerau%20Northland.pdf.

Treaty of Waitangi Act 1975. New Zealand Legislation, www.legislation.govt.nz/act/public/1975/0114/latest/DLM435306.html.

United Nations. "Renewable Energy – Powering a Safer Future." United Nations, 2025, www.un.org/en/climatechange/raising-ambition/renewable-energy.

Upton, John. "Solar Farms Threaten Birds." Scientific American, 27 Aug. 2014, www.scientificamerican.com/article/solar-farms-threaten-birds/.

Waitangi Tribunal. Ngawha Geothermal Resource Report (Wai 304). Brooker and Friend Ltd., 1993, www.nzgeothermal.org.nz/downloads/Ngawha-Geothermal-Resources-1993.pdf.

---. “Tribunal Releases Interim Report on Draft Geothermal Strategy.” Waitangi Tribunal, 28 Nov. 2025, waitangitribunal.govt.nz/en/news/tribunal-releases-interim-report-on-draft-geothermal-strategy.

Wilson, Andrew, et al. Kaitiakitanga and the Next 10,000 Megawatts. New Zealand Geothermal Association, 2015, www.nzgeothermal.org.nz/downloads/Kaitiakitanga-Guardianship-of-Geothermal-Wilson-et-al-2015.pdf.

---. Kaitiakitanga Future Geothermal Innovation and Direct Use. New Zealand Geothermal Association, 2017, www.nzgeothermal.org.nz/downloads/Maori-geothermal-development-opportunity-and-aspitrations-Wilson-et-al-2017.pdf.

Yong, Rachel, et al. A Deprivation and Demographic Profile of the Northland DHB. University of Auckland, 2017, www.fmhs.auckland.ac.nz/assets/fmhs/soph/epi/hgd/docs/dhbprofiles/Northland.pdf.

Physical and Psychological Challenges in Volleyball: Injury Prevention, Emotional Regulation, and Integrated Training Strategies By Andrew Huang

Introduction

Volleyball is a fast and physically demanding team sport. A player may need to jump, land, sprint, change direction, serve, pass, and communicate within only a few seconds. These actions require strength, endurance, body control, and technical skill. At the same time, volleyball also places pressure on athletes psychologically because players must make quick decisions, recover from mistakes, and stay calm during important points. For these reasons, volleyball performance should not be studied only as a physical issue. It is also connected to mental health, emotional regulation, and the way athletes manage stress during training and competition.

This research paper examines the physical and psychological challenges faced by volleyball athletes and explores practical strategies that may help them train more safely and perform more consistently. The main argument is that volleyball athletes need a balanced approach that protects the body, supports the mind, and connects both areas through purposeful training. The paper focuses on three areas: common volleyball injuries and prevention, mental factors such as anxiety and resilience, and the integration of physical and psychological routines.

Methodology

This research paper was conducted through secondary research. Instead of collecting new data from volleyball athletes, the paper analyzes existing information from academic journals, sports medicine resources, and research databases. The study compares different kinds of evidence, including injury studies, sports psychology research, and strength and conditioning literature.

First, literature on sports-related injuries was reviewed to understand how poor technique, insufficient warm-ups, overtraining, and lack of recovery may contribute to common volleyball injuries. Sources from sports medicine and injury-prevention research were used to identify frequent injury locations and prevention methods. This helped explain why volleyball athletes need strength, mobility, and safe movement patterns.

Second, research on sports psychology was examined to understand how mental factors influence athletic performance and well-being. This part focused on stress, anxiety, confidence, burnout, emotional regulation, and resilience. It also considered how coaches, sports psychologists, and structured mental skills training may help athletes manage pressure more effectively.

Third, articles on strength and conditioning, recovery, and integrated training were reviewed. These sources helped explain how purposeful training, recovery routines, proper equipment, sleep, nutrition, and psychological strategies can work together to support peak performance. By combining these areas, the paper develops recommendations for student athletes and school volleyball programs.

Results and Discussion

The first major finding is that physical health is essential because volleyball places repeated stress on the body. Volleyball requires explosive power, quick lateral movement, frequent jumping, and repeated landing. When athletes lack strength, endurance, or body awareness, they may become tired earlier and lose control of their movement. Fatigue can also affect technique. For example, a tired player may land with poor alignment, swing the shoulder with weaker control, or react slowly to a ball. Over time, these patterns can increase the risk of both acute injuries and overuse injuries.

Common injuries in volleyball include ankle sprains, finger sprains, anterior cruciate ligament injuries, patellar tendinitis, shoulder problems, low back pain, and knee pain. Bere et al. reported that match play created a higher time-loss injury risk than training. It shows that volleyball injuries are not random accidents only; they are also related to the intensity of competition, player position, movement demands, and repeated stress (1133). The FIVB Medical Commission similarly explains that volleyball is relatively safe compared with some sports, but players still face specific injury risks that coaches should understand (Reeser and Bahr 1).

Injury prevention is also connected to burnout. Burnout can be understood as physical and mental exhaustion that reduces motivation, enjoyment, and performance. When athletes train too much without enough recovery, they may feel tired, frustrated, or less confident. Physical fatigue can make an athlete more vulnerable to mistakes, while emotional fatigue can make training feel meaningless. A planned training structure is therefore important. Periodization, recovery days, mobility work, and gradual changes in training load can reduce unnecessary stress on the body and give athletes time to adapt (Sheppard et al. 259).

The second major finding is that volleyball depends heavily on decision-making under pressure. Volleyball is not only a game of physical skill. During a rally, players must decide where to serve, how to pass, whether to tip or attack, how to adjust a block, and how to communicate with teammates. A calm and focused player can process information faster and respond with better judgment. A nervous or distracted player may rush, overthink, or focus too much on the previous mistake. This is why mental skills such as focus, confidence, and emotional control are directly connected to technical execution.

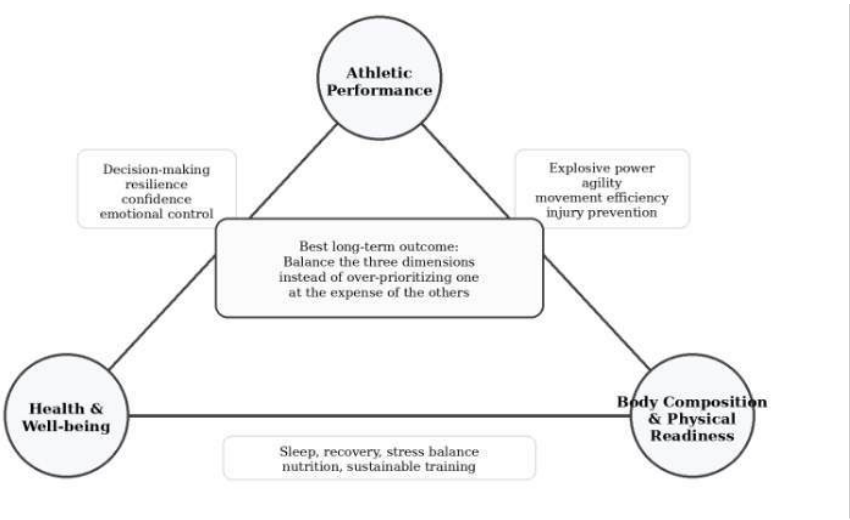


Figure 1. Conceptual Triangle for Sustainable Volleyball Development
Source: Author's own conceptual model

As shown in Figure 1, the triangle is a conceptual model that helps explain why strong performance depends not only on strength and skill, but also on recovery, emotional regulation, and healthy physical preparation. This broader view is consistent with the World Health Organization's understanding of health as more than simply the absence of disease. For this reason, the triangle shows that volleyball performance and well-being should be developed together, not treated as separate goals.

The third major finding is that resilience and adaptability are important for consistent performance. Volleyball matches often include sudden changes in momentum, controversial calls, missed serves, blocked attacks, and communication errors. A mentally resilient player is able to recover from these situations and return attention to the next point. In contrast, a player who keeps thinking about mistakes may become passive, lose confidence, or spread negative emotion to teammates. This matters especially in a team sport, where one player's emotional state can influence the rhythm and energy of the whole team.

Research by Vaccaro et al. compared elite volleyball players with amateur athletes and found that elite volleyball players reported lower depression and trait anxiety scores than amateur athletes (999). This shows that elite volleyball athletes have an overall stronger mentality than amateurs, which is a skill they have developed from training for years to compete at a high level. But this result should not be interpreted as proof that elite athletes never struggle mentally. Instead, it suggests that structured, supervised sport participation and long-term training may be related to stronger coping skills. Therefore, mental resilience can be trained and should be treated as part of athletic development, not as a separate personal trait.

Overall, the results show that physical and mental health are connected. A physically tired athlete may become more anxious or less focused. A mentally stressed athlete may move with tension, ignore recovery, or make poor decisions during a match. Therefore, volleyball training should not divide physical preparation and mental preparation completely. A more

useful model is to treat the athlete as a whole person whose body, mind, and environment influence one another.

Recommendations

The first recommendation is to use biomechanical and neuromuscular training to reduce injury risk and support better movement quality. Biomechanical training helps athletes understand how their bodies move during jumping, landing, cutting, serving, and attacking. Neuromuscular training can include landing drills, balance exercises, hip and glute strengthening, ankle stability work, core training, and controlled change-of-direction practice. These exercises are useful because they teach athletes to move safely under pressure instead of only relying on natural habits.

Although many injury-prevention programs are studied in other sports, the same principles are relevant to volleyball because volleyball also requires repeated jumping and landing. For example, athletes should learn to avoid knee collapse during landing, absorb force with the hips and knees, and maintain trunk control. These habits can reduce stress on the ankle, knee, and back while also improving movement efficiency. Coaches should include short injury-prevention routines in regular practices rather than treating them as extra work. This makes the recommendation more realistic for school teams with limited time and resources.

The second recommendation is to include emotional regulation training as part of volleyball practice. Emotional regulation helps athletes manage stress, stay focused, and make better decisions after mistakes. Gross's process model of emotion regulation explains that people can regulate emotions at different stages, such as by changing the situation, shifting attention, changing thoughts, or managing responses (282). In volleyball, this can be applied through routines such as breathing before a serve, using cue words after an error, focusing on the next point, and communicating with teammates in a calm way.

More structured psychological strategies can also be helpful. Cognitive behavioral therapy, or CBT, teaches people to identify automatic negative thoughts and replace them with more realistic and useful thinking patterns (American Psychological Association). For athletes, this may include changing thoughts such as "I always fail under pressure" into "I made one mistake, but I can reset and focus on the next ball." Hayes and Smith suggest that Acceptance and Commitment Therapy, or ACT, emphasizes accepting uncomfortable thoughts and feelings while still acting according to values such as effort, teamwork, and discipline. In a volleyball context, these ideas can help players perform even when they feel nervous.

The third recommendation is to build an integrated physical and mental health routine. This routine should include regular strength and conditioning, warm-ups, cool-downs, mobility work, sleep routines, hydration, balanced nutrition, and short mental skills exercises. The purpose is not to make training more complicated, but to make it more complete to support peak performance. For example, a practice session could begin with a movement-preparation warm-up, include technical drills with attention cues, and end with a short reflection on fatigue, confidence, and recovery needs.

Coaches can also use simple monitoring tools. Athletes may rate sleep quality, soreness, stress, and energy on a short weekly form. This type of tracking helps coaches notice when athletes are overtraining or struggling emotionally. It also encourages athletes to become more aware of their own body signals. For school teams, this can be done without expensive technology.

Conclusion

This paper examined the physical and psychological challenges that influence volleyball athletes' performance and well-being. The findings suggest that volleyball athletes face three connected problems: injury risk from repeated high-intensity movement, mental pressure during competition, and the difficulty of balancing physical training with psychological health. These problems are serious because they can affect not only match performance but also motivation, confidence, and long-term participation in sport.

The recommendations are to use biomechanical and neuromuscular training, include emotional regulation skills in regular practice, and build an integrated routine that connects physical preparation with mental health support. These strategies are practical for student athletes because they do not require a completely new training system. Instead, they improve what teams already do by making warm-ups, recovery, communication, and mental reset habits more intentional. Future research could include interviews or surveys with volleyball athletes to compare how players at different levels experience injury, stress, and recovery. This would provide more direct evidence from athletes.

Works Cited

American Psychological Association. “Cognitive Behavioral Therapy.” American Psychological Association, https://www.apa.org/ptsd-guideline/patients-and-families/cognitive-behavioral. Accessed 23 June 2026.

Bere, T., et al. “Injury Risk Is Low among World-Class Volleyball Players: Four-Year Data from the FIVB Injury Surveillance System.” British Journal of Sports Medicine, vol. 49, no. 17, 2015, pp. 1132–37. https://doi.org/10.1136/bjsports-2015-094959.

Gross, James J. “The Emerging Field of Emotion Regulation: An Integrative Review.” Review of General Psychology, vol. 2, no. 3, 1998, pp. 271–99. https://doi.org/10.1037/1089-2680.2.3.271.

Hayes, Steven C., and Spencer Smith. Get Out of Your Mind and into Your Life: The New Acceptance and Commitment Therapy. New Harbinger Publications, 2005.

Reeser, Jonathan C., and Roald Bahr. “Principles of Prevention and Treatment of Common Volleyball Injuries.” FIVB Medical Commission, https://www.fivb.com/wp-content/uploads/2024/03/FIVB_Medical_Injury_Prevention.pdf. Accessed 23 June 2026.

Sheppard, Jeremy, et al. “Volleyball.” Routledge Handbook of Strength and Conditioning: Sport-Specific Programming for High Performance, edited by Anthony Turner, Routledge, 2018, pp. 259–79.

Vaccaro, Maria G., et al. “Individuals’ Depression and Anxiety Might Be Influenced by the Level of Physical Activity and Expertise: A Pilot Study on Elite Volleyball Players and Amateur Athletes.” Sport Sciences for Health, vol. 17, 2021, pp. 999–1005. https://doi.org/10.1007/s11332-021-00767-2.

World Health Organization. “Constitution.” World Health Organization, https://www.who.int/about/governance/constitution. Accessed 23 June 2026.

The Influencer Impact: Social Media Influencers and Cosmetic Purchasing Behaviors and Habits Among Teenage Girls Ages 14–18 in Southern California By Grace Hirakawa

ABSTRACT

This study examines how social media influencers on platforms such as TikTok and Instagram impact the cosmetic purchasing behaviors of teenage girls aged 14–18 in Southern California. The central research question asks whether influencer content shapes consumer habits within this demographic. A correlational research study was conducted using a survey distributed to 62 participants, all of whom were teenage girls in Southern California. Participants were presented with a cosmetic product through three different stimulus conditions and asked to rank their likelihood of purchasing the product after viewing each stimulus. This product was then repeated with a second cosmetic product.

The findings indicate a clear relationship between the product promoting content of influencers and increased likelihood of purchasing cosmetic products. Products associated with influencers were more persuasive, suggesting that social media plays a significant role in shaping consumer behaviors. The results also reveal patterns of overconsumption, as participants often prioritized desired products over existing possessions. Overall, the study suggests that consumer behavior is strongly influenced by social media factors that encourage nonessential spending. These findings highlight the susceptibility of teenage consumers to influencer marketing and contribute to a broader understanding of how social media is shaping modern consumer culture and promoting increased consumption.

INTRODUCTION

Social media has become deeply embedded in the everyday lives of younger generations in the United States. From 2022 to 2023, the number of users increased by 150 million, representing a 3.2% year over year rise (Nyst, 2023). As its user base grows, social media increasingly influences decision making, functioning as both entertainment and a driver of conscious and subconscious choices, reflecting greater daily dependence on these platforms.

The rise in social media usage has led to the emergence of a new job, the influencer, to gain popularity. According to Coursea, an American online course provider created by a Stanford professor, a social media influencer is “a content creator who has gained a following on social media platforms by establishing credibility and expertise in a specific area, or niche” (Coursera, 2023). Like any job, influencers base the content they post on what creates the most revenue. An example of how influencers make money on social media is by working in partnership with established brands through paid partnerships, commissions, and affiliate marketing, encouraging followers to purchase featured products. In return, the more a product is sold, the more money the influencer receives. Because influencers’ incomes are frequently tied to engagement and sales, their content is often strategically designed to maximize consumer interest and spending.

To further elaborate, consumer behavior or consumer habits is defined by Michael R. Solomon in the book titled *Consumer Behavior: Buying, Having, and Being* as the processes

involved when individuals or groups select, purchase, or use products or services to satisfy needs and desires (Solomon, 2018), and consumerism is defined as "a belief that personal happiness and social status depend on the purchase of goods and services, often beyond basic needs," by the Encyclopedia Britannica (Duignan, 2026). As influencer marketing becomes more prevalent, it may contribute to a broader culture of consumerism, in which purchasing behavior extends beyond basic necessity and becomes tied to identity, status, and self-image.

Consumer behaviors that promote consumerism are vital part of the economy as proven through statistics from the United States Bureau of Economics Analysis which states that in the first quarter of 2025, consumption expenditure represented 67.7% of the nation's Gross Domestic Product or the total monetary value of all the final goods and services produced within a country's borders over a specific period (U.S. Bank, 2025). However, increased consumer spending and overconsumption are associated with several negative consequences, particularly environmental harm. Allison Denton, an associate advisor on banking at the Mayer Brown law firm, notes that the fashion and textile industry account for approximately 20% of global water waste and are mismanaged and they are the second-largest contributor to water pollution (Denton, 2023). These broader patterns highlight some of the wider impacts of rising consumerism, which can be viewed through the economic dependency of consumerism.

To add on, teenagers, with easily impressionable minds, fall trap to the consumer market due to influencers' promotion of products. One of the most prominent types of influencers are cosmetic or beauty influencers who promote various products relating to enhancing appearance or personal hygiene. Influencers regularly partner with brands within the cosmetic industry, which is defined by Rasha Mahmoud as "the business of making and selling products that help people with their appearance and personal hygiene" (Mahmoud, 2023), and both parties benefit from the video content of beauty influencers Correspondingly, their content encourages viewers to become consumers and purchase the products they are highlighting, even if said viewers own a single or multiple versions of the same product. Another example of how influencers partner with brands to promote their products is by explaining to viewers the uniqueness of the product and its necessity while also physically applying to the product in order to show why the product is so special. Influencers who are promoting the purchase of cosmetic products are in turn promoting consumerism and changing consumer behavior which have already been linked to negative impacts on the environment. If it can be proven that influencers have momentous impacts on the consumer habits of the cosmetic industry within teenage girls, then they will be more conscious about what they are endorsing and encouraging followers to not only view but also purchase.

LITERATURE REVIEW

The noticeable increase in consumerism has been recognized in multiple industries and fields. Allison Denton, financial advisor at Mayer Brown, analyzes fashion consumerism through both quantitative and qualitative methods, linking social media to buying habits within the fashion industry. She also examines how fast fashion, which became popular because of social media, impacts multiple aspects of society (Denton, 2023). Denton exemplifies through her

research that changes in the fashion industry harm the environment through chemical emission and water waste, exploit cheap labor, and violate intellectual property laws by copying original designs to make cheaper replicas. Additionally, she states that the harmful effects of the fashion industry will continue to grow in significance as the fashion industry itself grows. Similarly, Sulissya Nur Syafa'ati and Johny Alfian Khusyairi, doctoral students from Universitas Airlangga, Indonesia, conducted a qualitative study showing a positive link and between social media use and thrifting or purchasing pre-owned clothing (Syafa'ati, Khusyairi, 2024). The practice of purchasing pre-owned clothes, or thrifting, used to be heavily looked down upon. However, in recent years, social media has changed this perception and turned thrifting into a popular trend. Moreover, the researchers note that although thrifting is considered environmentally friendly, its rise on social media has encouraged overconsumption. The new trend illustrates how impressionable audiences can be and how social media driven consumerism can lead to negative consequences. Moreover, researches from the Taras Shevchenko National University of Kyiv, Olena Naumova and Mariia Naumova, study how fashion companies that falsely claim to be sustainable are not able to maintain their sustainability while also keeping up with supply and demand and keeping their cost of production low (Naumova, 2024). As a result, many fashion companies use deceptive practices to build trust and encourage purchasing. The manipulation extends beyond false sustainability claims to broader strategies that foster misleading trust and drive consumerism. The effects of social media on consumerism and industry deception have been widely studied in the fashion sector, but remain underexplored in other industries. In particular, there is limited research on how social media shapes consumer behavior within the cosmetic industry. Additionally, current studies fail to study broader influencer driven consumerism and instead focus on the effect of specific trends on consumerism.

In addition, the shift in entertainment from traditional forms such as the television and radio to social media has forced companies to develop new forms of marketing in order to promote sales. Among younger generations, the rise of social media has created new retail platforms, such as TikTok Shop and Facebook Marketplace, prompting companies to merge entertainment with e-commerce to attract and convert viewers into consumers. According to Xanayra Marin Lopez, brands are putting larger investments into marketing their products on social media platforms and this switch has been very beneficial for brands. Lopez also writes about how companies are also using social media influencers to market their products and in return giving the influencers commission on the products they push to their viewers (Marin-Lopez, 2025). The recent development of new marketing strategies allows brands to benefit from the societal shift to a digital world and stay relevant to consumers. Evidence of the positive role social media can play in brand marketing is seen through the research of I Made Sura Yoga Dinata, Tjokorda Gde Raka Sukawati, Gede Suparna, and I. Gst. Ayu Kt, four researchers for the Indonesian Journal of Advanced Research. In their case study of the failing Bali-based brand Kober Mie, a shift to social media focused marketing improved the brands image, increased repeat purchases, and allowed the brand to grow in success (Dinata, Sukawati,

Suparna, Giantari, 2025). The work of the researchers shows how social media marketing can be effective in helping brands and companies be successful in the changing society and grow into bigger companies with increased sales. The studies done in regards to brands utilizing social media explore how it can be beneficial for the brands; however, they do not acknowledge how social media can have a negative or positive effect on consumer habits and behaviors.

To add on, the studies and research done within the field and presented within the literature review focus on young adults and fail to acknowledge the effects of social media on the consumer habits of teenagers. Additionally, many studies focus on consumer habits in countries such as Indonesia, Bali, and Ukraine, but do not explore consumer habits within the United States.

Given existing gaps in the literature and ongoing concerns surrounding consumerism and social media, this study addresses the following research question: to what extent do social media influencers on platforms such as TikTok and Instagram impact the cosmetic purchasing behaviors and habits of teenage girls aged 14–18 in Southern California? It is hypothesized that social media influencers greatly impact the cosmetic consumer habits of teenage girls ages 14-18 in Southern California by convincing impressionable teenagers to buy products that they do not need nor truly want and thus promote overconsumption.

METHODOLOGY

Method of Inquiry

The method of inquiry that guided the research was a correlational research study which is defined by the Texas State University as a research study that examines the relationship between two or more variables to see if they are related (Kelemen, 2023). A study done by Filya Acikgoz and Sebnem Burnaz, a doctoral student and professor at the Istanbul Technical University, worked to find a relationship between influencer sponsored content on YouTube and the perceived value of sponsored content and viewers' attitudes towards YouTube influencers (Acikgoz, Burnaz, 2021). Thus, the two researchers conducted a correlational research study to successfully locate correlation between their variables. Looking for a relationship between teenage girls' consumer habits and influencers allowed for the examination of whether influencer marketing correlates to their purchasing behavior and desire for certain products and the degree of said correlation.

Primary Tool

Additionally, the primary tool used within the research was survey. The same study as mentioned previously by Acikgoz and Burnaz administered a survey to over 400 Turkish university students who were active YouTube users and each participant was asked questions that measured their views on sponsored YouTube content and influencers (Acikgoz, Burnaz, 2021). The usage of a survey allowed the researchers to obtain a large sample size of respondents due to the ability to easily disseminate the survey and allowed the researchers to use purposive

sampling, meaning all respondents had previous experience with YouTube. Similarly, a study done by N. Thagarajan and P. Sasikumar, researchers at the VELS Institute in Chennai, India, focusing on digital marketing and consumer behavior, used a survey to examine how Amazon's digital marketing strategies, influence consumers' trust, interest, and decisions to buy electronic products (Thiyagarajan and Sasikumar, 2025). Utilizing a survey allowed the researchers to get the opinions of over 100 participants who were not near them in geographical location but possessed expertise and experience that the two researchers were looking for, which again is purposive sampling. Within this research project, a survey was the chosen tool in order to get a large sample size and, through purposive sampling, obtain people who had experience with social media and the cosmetic industry, which was established through pre-survey questions.

Furthemore, keeping the entire survey digitized through usage of the Google Forms allowed for an easier way to share the survey with more people and ensure a sufficient amount of responses in order to draw valid conclusions about the correlation between the two variables. Additionally, Google Forms allowed the survey to be categorized into different sections which separated the Informed Consent Form, pre-survey questions, and the survey, maintaining overall organization. Moreover, the digitized element kept all responses confidential under a single Google account.

Background of Respondents

All respondents asked to take part in the survey were female due to the fact that the primary consumers of the cosmetic industry are females. A 2024 survey found that women buy cosmetics significantly more often than men, with men more likely to purchase only once a year or less (Statista, 2024). Furthermore, to fill the gap of younger age representation in the existing literature, the study focused on teenage girls aged 14–18. Age 14 was selected as the beginning of adolescence, when social media use and exposure to cosmetic products typically increase, while age 18 was used as the upper limit, marking the transition into adulthood.

Stimulus Material

In order to specifically focus on the extent to which social media has an impact on the consumer habits of viewers, stimuli were embedded within the survey to signify the various types of marketing and social media content brands use to promote products. The survey contained two sets of three stimuli, each containing a different cosmetic product. More specifically, mascara and lipgloss were chosen as the two stimulus products because a 2024 Statista report stated that lipgloss was the number one most used cosmetic product among teens and mascara was the most purchased eye cosmetic product among women (Statista, 2024). The familiarity of the two products meant participants were likely already aware of them and able to assess their consumer habits based on prior knowledge. Stimulus set one contained the Armani Beauty Vertigo Lift Mascara and stimulus set two contained the Givenchy Rose Perfecto Liquid Gloss and Balm. The two products were chosen because they were relatively new and not widely popular among teenagers, minimizing prior bias. Both stimulus sets one and two each contained three separate stimuli: the first one was a simple picture of the product, the second was a short

promotional video of just the product, and the third was a video of an influencer recommending the product by physically applying it and verbally explaining its qualities. The specific stimuli used are located within Appendix C.

Survey Content

The survey was sent to 100 potential respondents and received 62 responses back, with the other 38 potential respondents either declining to participate or failing to complete the survey. In addition, the survey itself began with the Informed Consent Form, which required a signature in order to be a participant, then the set of pre survey questions which were used to establish the experience and current consumer habits of the participants, and finally the survey questions. More specifically, each participant was shown the first stimulus from the first set of stimuli then asked to rank their likelihood to purchase on a scale from 1-10. This process was then repeated for each stimulus within the first set and then repeated for all the stimuli within the second set. Following this, all respondents were asked an open ended question: if you believe you would purchase this product, please explain what most inclined you to purchase it. Finally, at the end of the survey, all participants were thanked for their time and reminded of their right to withdraw their answers from the study at any given time.

Approach for Analyzing Survey Results

Subsequently, once all responses were submitted, the responses of participants were converted into quantitative and qualitative data that would be further analyzed and compared in order to find a correlation through the use of Google Sheets and automatically generated graphs from the Google Form tool.

FINDINGS

Pre-Survey Results

In order to establish prior experience of social media and the cosmetic industry, participants were asked a series of pre-survey questions. To begin, the participants were asked what social media platforms they used on a regular basis with the options being TikTok, Instagram, Snapchat, YouTube, Other, and none. 59 out of the 62 respondents selected one or more social media platforms and 3 stated that they did not use any forms of social media. The 59 participants who do use social media show the integration of social media within younger generations and also ensure that participants are already familiar with social media and social media influencers. The 3 that stated they used no form of social media were kept in the data set to get insight into how influencers might affect those who are not social media users. Furthermore, all participants were asked to select the types of cosmetic products they use on a regular basis with the there being 13 options to select from and an additional option to type in other products. The results showed that 100% of respondents used multiple types of cosmetic products on a daily basis thus establishing knowledge within the cosmetic industry and cosmetic products. Finally, within the pre-survey sections, all respondents were asked to select all the types of cosmetic products that they own two or more versions of and were given the same

options of 13 types of cosmetic products and an option to write in other products. The resulting data showed that 100% of all respondents owned multiple versions of multiple types of products, establishing previous consumer habits and a baseline for what participants already owned within their collection of cosmetic products. The full results of the cosmetic related pre-survey questions can be found in Appendix D

Stimulus Set One: Mascara

Following the Pre-Survey Questionnaire, all respondents were prompted with the first set of three stimuli materials, which contained the Armani Beauty Vertigo Lift Mascara and can be found in Appendix C. In between each stimulus, all participants were asked to rank their likelihood of purchasing the product on a scale from 1-10, with 1 being not likely and 10 being very likely. The results from the section of the survey are as follows and the full results extracted from the survey can be found in Appendix D.

Table 1. Most Common Responses to Stimulus Set One

Stimulus Source	Most Commonly Selected Answer	Second Most Commonly Selected Answer	Third Most Commonly Selected Answer
1.1	3	2	1 and 4
1.2	5	6	4
1.3	8	7	6

The table shows a shift in the top three most commonly selected answers as respondents reported being more likely to buy the product as the stimulus changed and more specifically participants reported being the most likely to purchase the product after seeing the third stimulus with the most common answers on a 1-10 scale being 8, 7, and 6. The numbers reported after seeing 1.3 are higher than the top numbers recorded after respondents viewed Stimulus 1.1 and 1.2. This trend can be observed by examining each column from top to bottom and identifying how the change in the stimulus was followed by a change in the likelihood of purchasing.

Stimulus Set Two: Lipgloss

Subsequently, the process was repeated with Stimulus Set Two, which was the Givenchy Rose Perfecto Liquid Gloss and Balm and can be found in Appendix C. Again, after being presented with each stimulus material, participants were asked to rate their likelihood of purchasing the product on a scale of 1-10. The full data set can be found in Appendix D and a portion of the data is presented below.

Table 2. Most Common Responses to Stimulus Set Two

Stimulus Source	Most Commonly Selected Answer	Second Most Commonly Selected Answer	Third Most Commonly Selected Answer

2.1	1	2	3 and 4
2.2	6	4	3
2.3	7	8	9

Similar to the findings from Stimulus Set One, the change of stimulus material presented resulted in an obvious increase in likelihood to purchase. As stated by the table, the most common answers after being shown Stimulus 2.1 were 1, 2, 3, and 4. Conversely, the most common answers after being shown Stimulus 2.2 were 3, 4, and 6 and after being shown Stimulus 2.3 were 7, 8, and 9, which shows an increase in the likelihood of purchasing as the stimulus material changes. Additionally, the table shows Stimulus 2.3 was responsible for having the highest numbers on a 1-10 scale of likelihood to purchase which correlates to the third stimulus fostering the highest sentiment of wanting to purchase the product and having a greater impact on consumer purchasing behavior.

Qualitative Analysis

Along with quantitative questions, participants were asked to provide qualitative data through short answer questions. For both the mascara and the lipgloss section of the survey, respondents were asked a single question: if you believe you would purchase this product, please explain what inclined you to purchase the product. Out of the 52 participants who filled out this section, every answer mentioned the third of stimulus or the influencer. More specifically, each respondent mentioned either being influenced by seeing the influencer apply the product, the way the influencer talked about the product, or the way the product looked on the influencer. Under the mascara section of the survey, one of the responses to the free response question was as follows.

> "The video (stimulus 3) had a real life demonstration of how the product worked. It also looked to be something that I would purchase since it was a mascara that made lashes look bigger and longer."

Additionally, in the lipgloss section of the survey, one participant stated that they would purchase the product and their reasoning why is as stated below.

> "I would purchase this product after seeing how well it looks on the influencer and hearing the positive factors the lipgloss has."

Another response in the same section that is similar to the one listed above came from another participant.

"After hearing the depth of what the girl said in the 3rd video I probably would buy this because she said it was a ph balancing lip gloss which I find interesting."

The responses presented encapsulate the trend of the data set as a whole as each response to the open ended question either explicitly mentioned the influencer, the girl in the video, or the third stimulus as being the decisive factor that most inclined them to want to purchase the product. Thus, the qualitative results show a common theme among all respondents in regards to the social media videos presented and its noticeable ability to impact consumer behavior.

DISCUSSION

Impacts of Social Media

In both the lipgloss and mascara stimulus sections of the survey, the results were fairly similar. When participants were asked to rank their likelihood to purchase a product on a scale from 1-10, Stimulus 1, which was a simple photo of the product, had the highest frequency of threes, twos, and one. Alternatively, Stimulus 3, which was the short video taken from social media of a woman applying the product while verbally recommending it, had the highest frequency sevens and eights on the scale of their likelihood of purchasing. The noticeable increase in likelihood of purchase can be traced to solely the change in media presented to respondents as all other variables including type of product and the brand of the product were kept constant in each section of the survey. Thus, social media influencers can be credited to having the largest impact on the increase of likelihood of purchasing compared to the other stimuli shown.

Relating back to research questions and hypotheses, a clear relationship or correlation is established between the content and promotion of social media influencers and the cosmetic consumer behaviors of teenage girls as the stimuli with the influencers made teenage girls the most likely to purchase the cosmetic products. Ultimately, the results indicate not only a correlation between the two, but also that influencers have an extensive impact on purchasing decisions and consumer habits.

Relation to Body of Knowledge

The study mentioned before by Nur Syafa'ati and Johny Khusyairi that examined the popular trend of thrifting clothes concluded that social media has played a large role in making the trend so popular which in turn made thrifting much more popular, interesting, and accepted within society (Syafa'ati, Khusyairi, 2024). Social media was found to not only reduce the stigma surrounding thrifting, but also encourage individuals to purchase thrifted clothing, thereby influencing consumer behavior. Additionally, the study highlights that this influence is partly driven by social media's role in promoting people to follow trends, motivating viewers to participate in trends such as the growing popularity of thrifting. The pattern of social media impacting consumer behavior is also seen within the cosmetic industry as shown by the research and data presented. In both fields, social media is clearly correlated with consumer behavior, as it

influences spending and shapes purchasing habits, thereby showing the impacts of social media are broader than a single industry. Although the two studies are similar in their overall conclusions, there are gaps within the body of knowledge that are filled due to this study. This study chooses to focus on teenage girls instead of young adults, it focuses on broader influencer driven consumerism rather than a specific trend, and narrows its scope to the cosmetic industry instead of the fashion industry.

Moreover, an additional study mentioned before by Allison Denton that studies how fast fashion has altered consumer purchasing behavior concludes that fashion consumption is strongly shaped by social trends and consumer pressure, and consumer habits become more impulsive and trend responsive (Denton, 2023). Overall, her research highlights how social media has impacted the popularity of the fast fashion industry which has then influenced consumer behavior and purchasing habits, thereby stating that social media has a correlation to consumer behavior. In comparison, both studies are successful at identifying a correlation between social media and consumer behavior, asserting that social media prominently promotes purchasing and overspending as people attempt to match cultural expectations and better assimilate into society. Their overlap in conclusions draws attention to the large and recognized impact of social media in the broader ideas of consumer behavior and that consumerism is being impacted in multiple different industries but lead to the same results. However, contrasts between the two lie in the gaps of Dentons's research such as social media's impact on industries other than the fashion industry and this study fills that gap by focusing solely on the cosmetic industry and how social media influencers are correlated to the consumer habits of shoppers within said industry.

Overconsumption

All respondents reported owning multiple versions of similar products, reflecting established consumer habits among teenage girls, including overconsumption, which is the purchasing of additional products despite already owning comparable items. To add on, although mascara and lip gloss were among the most common answers to have multiple versions of, 80% of participants rated their likelihood of purchasing mascara as 6 or higher, and 70% did so for lip gloss. This elevated likelihood of purchasing said products, despite existing ownership, suggests a disregard for what one already owns and a prioritization of wants over needs and establishing a shift in consumer purchasing behavior towards a value in overconsumption and excess spending. Overall, the findings indicated that exposure to consumer culture may contribute to continued purchasing behavior even when products are already owned.

CONCLUSION

The study found a correlation between consumer habits and influencer marketing. More specifically, the responses gathered from over sixty participants highlighted the common trend that products more heavily marketed by an influencer increase one's likelihood of purchasing the product to a significant extent, regardless of the two products being previously owned. There are evident gaps within the previous body of knowledge such as a lack of research on how social

media correlates to consumerism in teenagers and a higher focus on the fashion industry over the cosmetic industry that are filled through the study. For example, the survey solely focuses on how the teenage consumer is impacted through measuring each respondents experience and perception towards the products. Additionally, the stimuli disseminated to respondents are only cosmetic products, thus reinforcing a concentration on the cosmetic industry. Thus, the conclusions and new understandings drawn assert a response to the initial research question, proving a notable correlation between social media influencers and consumer habits

Implications on the Community

Following the conclusions drawn from the data and discussion, there is a clear correlation between social media influencers promoting cosmetic products and the consumer habits of teenage girls. The relationship has significant implications, as patterns of overconsumption developed during adolescence may carry into adulthood, potentially leading to poor financial habits such as impulsive spending and an inability to effectively save and manage money. Such behaviors can place individuals in financially unstable situations later in life. However, the study also highlights the role of social media in shaping these habits, encouraging greater awareness about how influencer content may encourage unnecessary purchases without critical evaluation. The findings also have broader implications for multiple stakeholders. Teenage girls may become more conscious of their susceptibility to influencer persuasion and the long-term effects of their spending habits. A more conscious and careful adoption of consumer habits will lead to a stronger financial compass and foster positive consumer habits that will carry into adulthood. Parents may be more likely to monitor both their children's social media exposure and purchasing behavior as they become more aware of how big of a role social media plays in consumer behavior. Furthermore, consumer industries such as the textile industry have horrendous effects on the environment, and by reducing the amount of products consumed, the industries responsible for destroying the environment will be forced to cut back on production, aiding society in healing from the negative implications that consumerism has caused.

Implications of the Field of Study

This study has important implications for the field of consumer behavior and social media research by reinforcing the role of influencer marketing in shaping purchasing habits among teenage audiences. The findings suggest that social media platforms are not only channels for product promotion but also powerful spaces where consumer attitudes and long term consumer behaviors are formed. Specifically, the correlation between influencer content and changes in consumer purchasing behavior highlights the need for further research into how digital marketing strategies affect youth decision-making and why the correlation between the two variables occurs in the first place. Additionally, this study contributes to a broader understanding of how early exposure to persuasive online content may influence financial literacy and spending patterns into adulthood. Overall, the results emphasize the importance of examining social media as a key factor in modern consumer behavior models.

Limitations

Although the method of the study aimed to gather data that was both credible and unbiased in order to draw larger conclusions, there were still limitations within the study. For starters, most respondents were located within the same school and only three were from an outside neighboring school. Although the geographical spread of the sample size is not as diverse as it could be, each respondent has their own personal experiences which are far more significant to the study than the location of the respondents. Concurrently, the topic at hand focuses on social media and the cosmetic industry, two elements that are not geographically bound to a certain area and instead available almost everywhere in the country and thus affect people similarly in various regions.

Moreover, another limitation to the study is pre-existing biases toward the brands of products presented in the stimulus or bias toward the influencers within stimulus three. Although bias is inevitable, it is not a factor that discredits or negates the study. The influencers selected both have a small but noticeable following, making it unlikely that teenagers have a bias toward them, and the products selected were reported to have both low amounts of total purchases on the Sephora Website. Additionally, connections to certain influencers or brands can be part of the reason why influencer promotions can be so effective in convincing teenagers to purchase products. Bias on content within the stimulus can further prove impact on consumer habits as teenagers are more trustworthy of influencers and brands they recognize and trust, thus continuing to foster consumer habits.

Future Research

The research conducted has a multitude of possibilities for future research beyond what is already studied. To begin, researchers could study how social media and influencers affect consumer habits in teenage boys instead of girls. Furthermore, the research conducted in both genders does not have to solely be focused on the cosmetic industry, but can examine other underrepresented industries such as sports and entertainment. Equally significant, another direction for future research could be increasing the age range to women in their early twenties all the way to older women who are past fifty. The older age range would establish how consumer habits and influenceability can differ with age and if the consumer habits women learn as teenagers is something carried through adulthood or refined with age. Additionally, a study into older age groups of women allows for a deeper understanding of how social media impacts teenagers as older women did not have access to social media during their youth. Overall, the possibilities for future research all allow for a deeper and more well rounded understanding of how social media, which has become deeply integrated within society, can affect consumer purchasing behavior, and how it may continue to shape purchasing decisions, financial habits, and cultural norms in increasingly significant ways.

Works Cited

Acikgoz, Fulya, and Sebnem Burnaz. 2021. “The Influence of ‘Influencer Marketing’ on YouTube Influencers.” International Journal of Internet Marketing and Advertising 15(2):201. doi: https://doi.org/10.1504/IJIMA.2021.114331.

Adams, Maxine. 2024. “Maxine Adams on Instagram: ‘Pretty Lip Balm Alert 🎀🌹 Ft. @Givenchybeauty Rose Perfecto Liquid Lip Balm - 210 PR* @Blackandwhitecomms #Givenchybeauty #Pinklips #Pinkmakeup #Prettymakeup #Ugc #Ugccreator #Beautyfavorites #Beautyhaul #Viralbeauty #Viral products #Vanitymakeup #Makeupunboxing #Makeupnews #Itgirl #Thatgirlaesthetic #Cleangirlaesthetic.’” Instagram. Retrieved March 27, 2026 (https://www.instagram.com/reel/DBteWlrIgsI/?igsh=NTc4MTIwNjQ2YQ==).

Armani Beauty. 2017. “Armani Beauty on Instagram: ‘in the Blink of an Eye. With Its High-Tech Formula and Infinity Loop Brush, the New VERTIGO LIFT MASCARA Provides a Sculptural Lift, Separating and Curving Each Lash for a 180-Degree Spreadout Look. #Armanibeauty #ArmaniVertigoMascara #LookUpForVertigo #Mascara #Makeup.’” Instagram.

Denton, Allison. 2023. “The Cost of Looking Good: How Fashion and Trend-Based Consumerism Impact the Economy, Law, and Environment.” Indiana Journal of Global Legal Studies 30(2):363–89.

Dinata, I. Made Surya Yoga, Tjokorda Gde Raka Sukawati, Gede Suparna, and I. Gst. Ayu Kt. Giantari. 2025. “The Role of Brand Image in Mediating the Influence of Social Media Marketing on Repeat Purchase of Kober Mie Denpasar Customers.” Indonesian Journal of Advanced Research 4(7):1397–1410. doi: https://doi.org/10.55927/ijar.v4i7.14973.

Duignan, Brian. “Consumerism.” Encyclopedia Britannica, 23 Apr. 2026, www.britannica.com/money/consumerism?utm_source

Gagliano, Siena. 2025. “Giorgio Armani’s Vertigo Lift Mascara Convinced Me to Break up with My Go-to Formula for 10 Years.” Marie Claire. Retrieved March 27, 2026 (https://www.marieclaire.com/beauty/giorgio-armani-vertical-lift-mascara-review/).

Givenchy Beauty. 2023. “ROSE PERFECTO TINTED LIQUID LIP BALM.” Givenchybeauty.com. Retrieved (https://www.givenchybeauty.com/us/p/rose-perfecto-tinted-liquid-lip-balm-F20100155.html).

Leichtnam, Annaleza. 2025. “Part 25 I Holiday Wish List Starts with This Set: Givenchy Rose Perfecto Hydrating Tinted Lip Balm & Gloss.” Tiktok.com. Retrieved March 27, 2026 (https://www.tiktok.com/t/ZP8bbPGV1/).

Kelemen, William, et al. “Correlational Research.” Research Methods in Psychology, Kwantlen Polytechnic University, 14 Jan. 2023, pressbooks.txst.edu/3402kelemen/chapter/correlational-research/.

Marin-Lopez, Xanayra. 2025. “How TikTok Shop Is Changing the Way Brands Reach Customers.” Retail Dive. Retrieved (https://www.retaildive.com/news/how-tiktok-shop-is-changing-the-way-brands-reach-customers/736726/)

Mahmoud, Rasha. 2023. “Beauty Industry: Overview, Players & Distribution.” Retail Dogma. Retrieved (https://www.retaildogma.com/beauty-industry/).

Naumova, Olena, and Mariia Naumova. 2024. “Sustainable Development and Greenwashing: An Analysis of Sufficient Consumption in the Fashion Industry.” Economics, Finance and Management Review 17(1):16–24. doi: https://doi.org/10.36690/2674-5208-2024-1-16

Nyst, Annabelle. 2023. “134 Social Media Statistics You Need to Know for 2023.” Search Engine Journal. Retrieved (https://www.searchenginejournal.com/social-media-statistics/480507/).

Nicole, Samantha. 2025. “Trying the New @Armani Beauty Vertigo Lift Mascara 🖤 the Results Were Absolutely Unreal.” Tiktok. Retrieved (https://www.tiktok.com/t/ZP8gcadys/).

Statista Research Department. “China: Frequency of Purchasing Cosmetics by Gender 2024| Statista.” Statista, Sept. 2024, www.statista.com/statistics/1552841/china-frequency-of-purchasing-cosmetic-products-by-gender/?utm_source=). Accessed 30 Apr. 2026.

Statista Research Department. 2024. “U.S.: Most Used Color Cosmetics Items among Pre-Teens & Teens 2024| Statista.” Statista. Retrieved March 27, 2026 (https://www.statista.com/statistics/1500252/most-used-color-cosmetics-items-among-pre-teens-and-teens-in-the-us/?utm_source=).

Solomon, Michael R., and Internet Archive. Consumer Behavior: Buying, Having, and Being. Internet Archive, 12th ed., Harlow, England : Pearson, 2018, archive.org/details/consumerbehavior0000solo_n4t7.

Syafa’ati, Sulissya Nur, and Johny Alfian Khusyairi. 2024. “Thrift Fashion Trends as a New Culture of Consumerism.” Airlangga Development Journal 8(2):134–45. doi: https://doi.org/10.20473/adj.v8i2.44267.

Thiyagarajan, N., and P. Sasikumar. 2025. “Digital Marketing Strategies and Their Effect on Consumer Behaviour toward Electronic Products on Amazon.” Journal of Neonatal Surgery 14(23s):1126–31. https://www.jneonatalsurg.com/

U.S. Bank. 2024. “Consumer Spending | U.S. Bank.” Www.usbank.com. Retrieved (https://www.usbank.com/investing/financial-perspectives/market-news/consumer-spending.html).

A Coin Flip from Twelve Yards? The Penalty-Shootout "First-Kicker Advantage" at the World Cup, and Where the Pressure Really Bites By Robert Sultan

Abstract

In a widely cited 2010 paper in the *American Economic Review*, José Apesteguia and Ignacio Palacios-Huerta reported that the penalty shootout is unfair: across 129 shootouts the team kicking first won about 60.5% of the time, an edge they blamed on the pressure of always kicking second, a claim influential enough that football's lawmakers trialed an "ABBA" kicking order to cancel it. I test it on every men's FIFA World Cup shootout (all 39, 1982–2026) and, going deeper, on the 360 individual kicks inside them. The advantage does not appear: the first-kicking team won just 17 of 39 (43.6%, $z = -0.80$, $p = 0.42$), and all four shootouts added in 2026 were won by the team kicking second. The World Cup alone is underpowered (only about 25% power to detect a true 60.5% edge), but it agrees with the two largest studies ever run (540 and 7,116 shootouts), which pool with it to 50.4%. The mechanism the theory needs is also missing: the second (chasing) team converts no worse than the first (70.5% vs 66.8%, $p = 0.45$). Pressure is real, but it comes from the stakes, not the order: on must-score-or-be-eliminated kicks conversion collapses to 50% against 70% otherwise ($z = -2.12$, $p = 0.034$), while even the higher-ranked side wins only half its shootouts (15 of 31). The statistical morals outlast the football: replication, power, and the gap between a real effect and the wrong explanation for it.

Keywords: penalty shootout, first-mover advantage, one- and two-proportion z-tests, statistical power, replication, weighted averaging, confidence intervals, FIFA World Cup

Introduction

The penalty shootout is football's most dramatic tiebreaker and its most criticized. A recurring complaint is that it is not even fair between the two teams: the side that kicks first is said to enjoy a real advantage, because the side kicking second is forever chasing, always taking its kick under the scoreboard pressure of "score or go out." This is not merely a fan's grumble. In an influential paper, Apesteguia and Palacios-Huerta treated the pre-shootout coin toss as a randomized natural experiment and reported that the first-kicking team won 60.5% of shootouts, a roughly 21-percentage-point edge attributed to psychological pressure (Apesteguia and Palacios-Huerta). The finding reached popular audiences and policy: football's lawmakers (IFAB) tried an alternating "ABBA" kicking order, modeled on the tennis tiebreak, specifically to cancel the supposed first-mover advantage.

A headline percentage from one dataset, though, is exactly the kind of claim a statistics course teaches you to re-examine, and on more than one level. This paper works on three. First, it tests the claim itself at the World Cup and asks whether 39 shootouts can even settle it. Second, it opens up the shootouts and looks at the 360 individual kicks inside them to test the *mechanism*: if the second team really chokes, its kickers should convert at a lower rate. Third, it asks where pressure *does* show up, by comparing kicks taken under different stakes. A closing question steps

back from pressure entirely to ask whether the better team (the higher-ranked side) wins the shootout at all. Five concrete questions in all:

The first kick. Does the team that kicks first win more than half of World Cup shootouts?

The sample. If it does not, is that because there is no advantage, or because 39 shootouts is too few to tell, and what do larger studies say?

The mechanism. Do kickers on the second (chasing) team convert at a lower rate than the first team, as the pressure theory requires?

The stakes. Do "must-score-or-lose" kicks convert worse than ordinary ones, and do "score-to-win" kicks convert better?

The favorite. Does the higher-ranked team win the shootout, or does the shootout erase that edge?

The claim is not unstudied, and the literature has moved against it. The original 60.5% came from 129 shootouts (Apesteguia and Palacios-Huerta). A larger replication of 540 shootouts found no statistically significant first-mover advantage (Kocher et al.), and a recent study of 7,116 shootouts, the largest ever assembled, found the first kicker within 1.8 percentage points of an even split (Pipke). FIFA's ABBA trial was quietly dropped in 2018. Against that backdrop the World Cup is a useful special case: a small but famous, carefully recorded set of shootouts where the folklore is loudest and every kick is on record: an ideal place to test both the claim and the story behind it.

Data and Methods

Two datasets

Dataset A: shootout level. I compiled every penalty shootout in the history of the men's FIFA World Cup: 39 shootouts across the 1982 through 2026 tournaments (no World Cup match was decided by penalties before 1982, and no third-place match has ever gone to a shootout). The 2026 tournament, the first with 48 teams and an added round of 32, contributed four new shootouts. For each I recorded the round, the two teams, the shootout score, the winner, and the hardest field to pin down: which team took the first kick, since the coin-toss winner chooses whether to kick first. Kick order was taken from kick-by-kick records, not inferred from the scoreline; this matters, because casual summaries routinely list the winner first and thereby mis-assign the opening kick. For shootouts from 1994 onward I also recorded each team's FIFA World Ranking at the time; the ranking did not exist before December 1992, so the eight shootouts of 1982–1990 have no ranking and are excluded from the ranking test. The full 39-row dataset is in Appendix B.

Dataset B: kick level. From the same records I reconstructed all 360 individual penalty kicks, in the exact order taken, each labeled by the kicking team, whether it was scored, and the running score before it. This is the dataset for the mechanism and stakes questions. Because a

shootout stops the moment it is mathematically decided, these 360 are every kick actually taken, not a fixed 10 per shootout.

Statistical methods

Most questions are coin-flip tests, so they use a one-proportion z-test against H_0: $p = 0.5$, with $z = (\hat{p} - 0.5) / \sqrt{(0.25/n)}$, a two-sided p-value, and a 95% confidence interval. For the headline first-kicker rate I build the interval two ways (the ordinary Wald interval and the Wilson score interval) to show the conclusion does not depend on which one I use. To compare two conversion rates (first vs. second team; high-pressure vs. ordinary kicks) I use a two-proportion z-test, and to combine the World Cup with the published studies I take a weighted average of their win rates, giving each study a weight proportional to its number of shootouts.

Two definitions carry the stakes analysis. A kick is a must-score kick if missing it would end the shootout in defeat (elimination): formally, if the kicker's team, having missed, could no longer catch the opponent's current score given its remaining kicks. A kick is a to-win kick if scoring it ends the shootout as a win. Both are computed directly from the running score and the kicks remaining, in code (Appendix A).

The distinctive step is a power calculation. A non-significant result only means "no effect" if the test had a real chance of catching one. So I ask: if the true first-kicker win rate were the celebrated 0.605, what is the probability that 39 shootouts would return a significant result at $\alpha = 0.05$? This turns "we found nothing" into the sharper "here is what we could and could not have found."

All tests use $\alpha = 0.05$. These are observational data: only knockout ties still level after extra time reach a shootout, and the 39 are the whole population of World Cup shootouts rather than a random sample. So I read the p-values as a description of how surprising each pattern is. Computation was done in Python with pandas, NumPy, and SciPy; code is in Appendix A.

Results

The first kick: a coin flip

At the World Cup, kicking first confers no visible advantage. Of the 39 shootouts, the team that kicked first won 17 and lost 22, a win rate of 43.6%. Against the coin-flip null this gives $z = -0.80$ and $p = 0.42$: nowhere near significant, and if anything below an even split rather than above it. The four shootouts added in 2026 sharpen the point, because the team kicking first lost all four. And the conclusion does not depend on how the interval is built (Table 1): both methods center the rate near 0.44 and stretch it across a band roughly from the high 0.20s to 0.59, straddling a coin flip.

Table 1. The World Cup first-kicker win rate (17/39), by two interval methods. Both agree: centered near a coin flip, and wide.

Method	Point estimate	95% interval
Ordinary (Wald) z-interval	0.436	[0.280, 0.592]
Wilson score interval	0.436	[0.293, 0.590]

Why one tournament cannot settle it: power and the bigger picture

That wide interval is the point of this section. It runs from about 0.28 to 0.59, so it is consistent with everything from a sizable first-kicker disadvantage through a coin flip to a modest advantage; 39 shootouts simply cannot separate those stories. A power calculation makes the limitation precise: if the true rate really were the celebrated 0.605, a sample of 39 would return a significant result only about 25% of the time. Failing to find the effect here is almost expected even in a world where the effect is real, so the World Cup, for all its fame, is too small to be evidence about the claim on its own.

To settle a question this size you need many more shootouts, and larger studies exist. Figure 1 and Table 2 place the World Cup beside the three published estimates and their pooled average. The pattern is what you expect from a claim that does not survive scrutiny: the original small-sample study found 60.5%, but as the sample grows the estimate slides toward 50% and its interval tightens around it. Averaging the four rates, each weighted by its number of shootouts, gives a pooled 50.4% (95% CI [49.3%, 51.5%]), a coin flip. The one study that stands apart is the original small sample; the enormous modern samples pull the pooled estimate firmly to 50%.

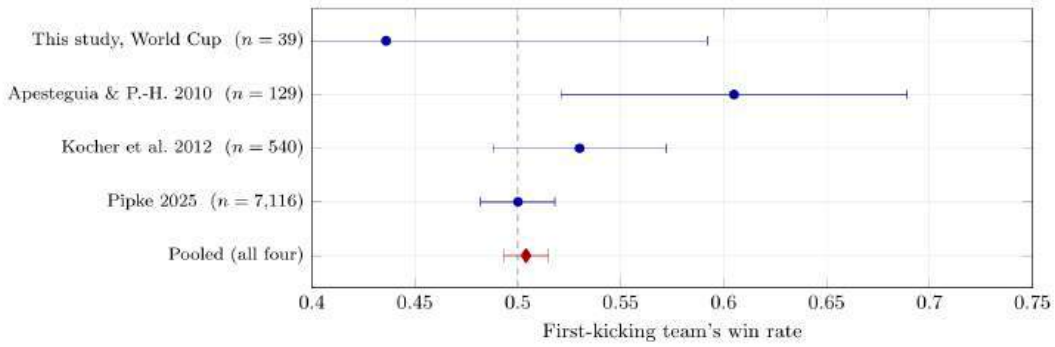


Figure 1. The first-kicker win rate across four studies and their size-weighted pooled average (red diamond), with 95% confidence intervals (for Pipke, the reported ±1.8-point equivalence bound). The dashed line marks a fair coin, 0.50. As the sample grows the estimate converges on 0.50 and its interval shrinks; the original 60.5% is the outlier, and the pooled estimate is 0.504. The World Cup (n = 39) agrees with the large samples but has an interval too wide to decide anything on its own. Kocher et al.'s point estimate is approximate; their reported result is "no significant advantage."

Table 2. The first-kicker advantage across studies. Larger samples converge on a coin flip.

Study	Shootouts (n)	First-kicker win rate	Verdict
Apesteguia and Palacios-Huerta (2010)	129	60.5%	Significant advantage
Kocher et al. (2012)	540	≈53%	Not significant
Pipke (2025)	7,116	≈50%	No advantage
This study (World Cup)	39	43.6%	Not significant
Pooled (weighted average)	7,824	50.4%	Coin flip

The mechanism, kick by kick: the chasing team does not choke

If the first-kicker advantage were real, it should leave a fingerprint on individual kicks: the second (chasing) team, taking every kick under "must-respond" pressure, ought to convert at a lower rate. It does not. Across all 360 kicks, players scored 68.6%, in line with the 70–75% usually quoted for shootouts. Split by kicking order, the two teams are almost identical: the first-kicking team's players converted 125 of 187 (66.8%), the second-kicking team's 122 of 173 (70.5%). If anything, the chasing team converts 3.7 percentage points higher (the opposite of the predicted direction), and the gap is not significant (two-proportion $z = -0.75$, $p = 0.45$). Whatever might make kicking a second harder, it does not lower the conversion rate, so the mechanism the theory needs is simply absent from the data.

The stakes, not the order: where pressure really bites

Pressure does bite, but not from the running order. The revealing split is by what is riding on the kick (Figure 2). On must-score kicks, where a miss ends the shootout in elimination, conversion collapses to 50.0% (13 of 26). On every other kick it is 70.1% (234 of 334). That 20-percentage-point gap is both large and statistically significant (two-proportion $z = -2.12$, $p = 0.034$), the clearest effect in the paper. The mirror image holds too: on to-win kicks, where scoring wins the shootout outright, players converted 26 of 30 (86.7%). The pressure to avoid losing depresses performance; the chance to seal a win does not. It is the asymmetry a psychologist would predict, and it is invisible to the first-kicker theory, which looks only at order.

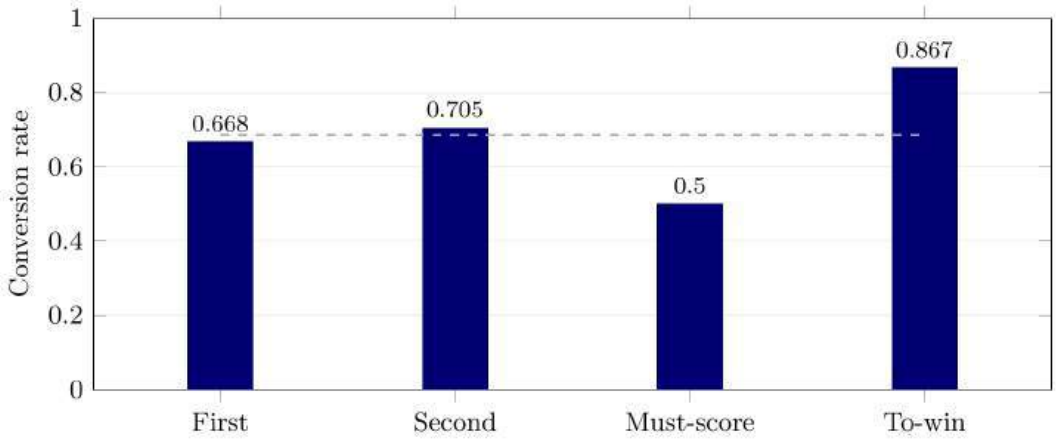


Figure 2. Penalty conversion by kick type, 360 World Cup shootout kicks. The first- and second-kicking teams (by order) convert almost identically and sit near the overall rate (dashed line, 0.686). Splitting by stakes tells a different story: "must-score-or-out" kicks crater to 0.500 while "score-to-win" kicks reach 0.867. Pressure is about what is at stake, not who kicks first.

The effect also builds as a shootout wears on (Figure 3 and Table 3): conversion holds around 72–73% through the first three rounds, then drops into the low 60s by the fourth and fifth, and the handful of sudden-death kicks split evenly. This is consistent with the stakes result (later

kicks are more often decisive), though the per-round samples are small and this is offered as a description, not a test.

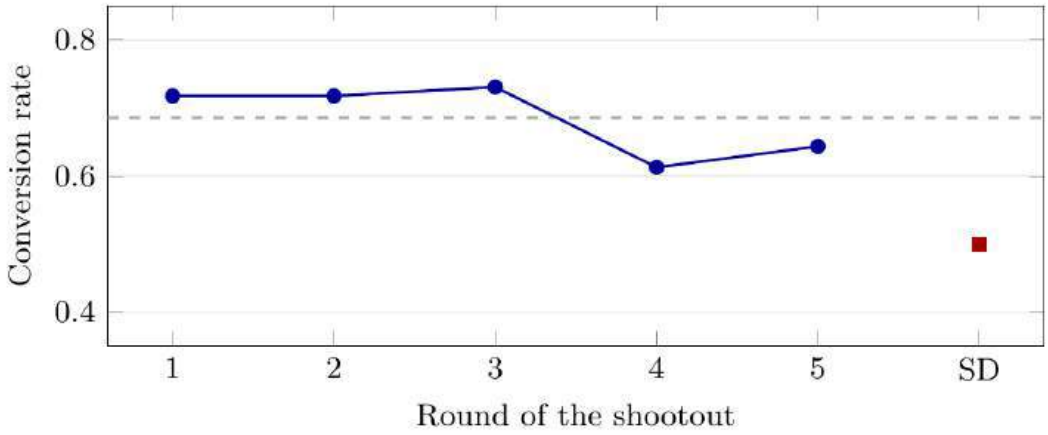


Figure 3. Conversion rate by round (dashed line = overall 0.686; sudden death in red). It holds near ~72% through round three, dips into the low 60s by rounds four and five, then the six sudden-death kicks split 3–3. Two forces plausibly overlap: the later a kick, the more often it carries elimination or victory stakes (see the stakes analysis above), and fatigue and accumulated tension may weigh on the takers. With so few late-round kicks, though, this is a suggestive pattern rather than a tested one.

Table 3. Conversion rate by round of the shootout (round 6+ = sudden death). Small per-round samples; descriptive only.

Round	**1**	**2**	**3**	**4**	**5**	**Sudden death**
Conversion	0.718	0.718	0.731	0.613	0.644	0.500
Kicks (n)	78	78	78	75	45	6

The favorite: the shootout erases quality

Finally, if neither the first kick nor the second team's nerves decide a shootout, is it the better team? Among the 31 shootouts in which both sides had a FIFA ranking (1994 onward), the higher-ranked team won just 15, a win rate of 48.4% ($z = -0.18$, $p = 0.86$), again a coin flip; 2026 fit the pattern, with the higher-ranked side losing three of the tournament's four shootouts. Reaching a shootout appears to wipe out the edge a team carried into the match. Spain is the clearest case: a perennial favorite, it lost World Cup shootouts to South Korea, Russia, and Morocco, each ranked far below it, and in 2022 failed to score a single penalty against Morocco. Croatia, by contrast, won all four of its shootouts across 2018 and 2022, including one against top-ranked Brazil. Once the game is level and the ball is on the spot, the form guide stops predicting the result.

Discussion and Limitations

The picture that emerges is coherent. The celebrated first-kicker advantage does not appear at the World Cup (43.6%); the two largest studies say it is not there in general; and, most

tellingly, the mechanism it depends on is missing from the individual kicks, since the chasing team does not convert worse. What is real is a different, larger effect the order-based theory never looks at: kicks taken to avoid elimination are converted far less often than ordinary kicks, and kicks taken to win are converted far more often. Pressure lives in the stakes, not the sequence. Several limits bound these claims.

• **Small sample / power (the shootout-level tests).** 39 shootouts have only about a 25% chance of detecting even a real 60.5% advantage, so the first-kicker result is best read as "consistent with a coin flip and with the large-sample consensus," not as standalone proof. The kick-level tests are on firmer ground (360 kicks, and the pressure effect clears significance), but a single team's record (Spain's 1–4, Croatia's 4–0) is an anecdote, not a test.

• **Defining "must-score."** The stakes classification is computed from the running score by a fixed rule (Appendix A); a different reasonable rule would reclassify a few borderline kicks. The effect is large enough (50% vs. 70%) that it would survive small changes, but the exact p-value is not sacred.

• **Ranking is an imperfect proxy for quality**, and under the pre-2018 formula occasionally tied teams; "the favorite" means "the higher-ranked side," which is slightly narrower than "the better team."

• **Observational data and selection.** Only knockout matches level after extra time reach a shootout, so these are a self-selected, very evenly matched set of games, which may itself be why the favorite's edge vanishes. Nothing here is a controlled experiment.

• **Data provenance.** Kick order and outcomes were reconstructed from public match records and cross-checked; a few miss types (save vs. off-target) carry minor uncertainty but never affect whether a kick was scored. One first-kicker label in the original summary (2002, Spain–Ireland) was corrected here: the opening kick was Ireland's, not Spain's. That is exactly the kind of error the kick-by-kick audit is meant to catch. The literature figures marked "≈" come from secondary summaries; the load-bearing facts (n = 129 at 60.5%, n = 540 not significant, n = 7,116 near 50%) are from the studies themselves.

It is worth being fair to the original study. Apesteguia and Palacios-Huerta reported a genuine, statistically significant result in their sample; the issue is not error or bad faith but generalization. A real-looking effect in 129 shootouts thinned toward nothing as the sample grew, in the ordinary way that sampling variability and publication combine to inflate a first finding.

The ABBA experiment is the institutional version of the same arc. Persuaded that kicking second was a real handicap, IFAB approved in 2017 a trial of an alternating order borrowed from the tennis tiebreak (team A kicks, then B twice, then A twice, and so on: ABBA rather than ABAB), so that neither side would repeatedly kick under "must-respond" pressure. It was used at the 2017 UEFA Under-17 and Women's Under-17 finals and by the English FA in the 2017 Community Shield and League Cup. Then, at its 133rd Annual Business Meeting in November 2018, IFAB discontinued the trial, citing limited support and the complexity of administering it. The reversal fits this paper's evidence: a plausible fix was built for a problem that the mechanism

data and the large samples suggest was never really there. The pressure ABBA was designed to neutralize is real, but, as the stakes analysis shows, it attaches to the stakes of a kick, not to whether a team kicks first or second, and no reordering of the sequence can legislate that away.

Conclusion

Tested against the whole history of World Cup shootouts, and then against the 360 kicks inside them, the penalty shootout looks like what players have long suspected: close to a coin flip, and cruel in a way that has nothing to do with who goes first. The first-kicking team won 43.6% of the time and the higher-ranked team 48.4%, both indistinguishable from chance, both in line with the largest studies ever run, even as they cut against a famous published claim. The mechanism that claim rests on is absent: the chasing team does not choke. What does move the numbers is the weight of the moment: kicks that must be scored to survive are missed more than half again as often as ordinary ones. The statistical morals outlast the football: a single significant result, however widely repeated, is a hypothesis awaiting replication; a small sample that finds nothing has not disproved an effect, only shown how little it could see; and confirming that an effect is real is a different task from confirming why. Table 4 collects the five questions and their answers.

Table 4. The five questions and what the data (all 39 men's World Cup shootouts and the 360 kicks inside them, 1982–2026) say about each.

Question	Answer	What the data show
Does the team that kicks first win more?	No	17/39 = 43.6% ($z = -0.80$, $p = 0.42$); a coin flip.
Can 39 shootouts settle it?	No	Only ~25% power to detect a 60.5% edge; the four studies pool to 50.4%.
Does the chasing (second) team choke?	No	Converts 70.5% vs the first team's 66.8% ($p = 0.45$); the mechanism is absent.
Do the stakes of a kick matter?	Yes	Must-score kicks convert 50.0% vs 70.1% otherwise ($z = -2.12$, $p = 0.034$).
Does the higher-ranked team win?	No	15/31 = 48.4% ($p = 0.86$); the shootout erases quality.

Works Cited

Apesteguia, José, and Ignacio Palacios-Huerta. "Psychological Pressure in Competitive Environments: Evidence from a Randomized Natural Experiment." American Economic Review, vol. 100, no. 5, 2010, pp. 2548–64. https://doi.org/10.1257/aer.100.5.2548.

International Football Association Board. "Decisions of the 133rd Annual Business Meeting: Discontinuation of the 'ABBA' Penalty Shoot-Out System Trial." International Football Association Board, Nov. 2018, Glasgow.

Kocher, M. G., et al. "Psychological Pressure in Competitive Environments: New Evidence from Randomized Natural Experiments." Management Science, vol. 58, no. 8, 2012, pp. 1585–91. https://doi.org/10.1287/mnsc.1120.1516.

"List of FIFA World Cup Penalty Shoot-Outs." Wikipedia, Wikimedia Foundation, 2023, https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_penalty_shoot-outs.

Pipke, D. "No Evidence of First-Mover Advantage in a Large Sample of Penalty Shootouts." Journal of Economic Psychology, vol. 108, 2025, article 102816. https://doi.org/10.1016/j.joep.2025.102816.

"2026 FIFA World Cup Knockout Stage." Wikipedia, Wikimedia Foundation, 2026, https://en.wikipedia.org/wiki/2026_FIFA_World_Cup_knockout_stage.

Appendix A: Reproducibility — Analysis Code

Two CSV files and two Python scripts reproduce every number. worldcup_shootouts.csv (39 rows) drives the shootout-level tests; worldcup_shootout_kicks.csv (360 rows, generated by the second script) drives the kick-level tests.

A.1 Shootout level: first-kicker, power, and the favorite

```
import pandas as pd, numpy as np
from scipy import stats

df = pd.read_csv("worldcup_shootouts.csv")

def one_prop_z(wins, n, p0=0.5):
    phat = wins / n
    z = (phat - p0) / np.sqrt(p0 * (1 - p0) / n)
    return phat, z, 2 * (1 - stats.norm.cdf(abs(z)))

first_won = (df["winner"] == df["team_first"]).sum()
print(one_prop_z(first_won, len(df)))            # 17/39, z=-0.80, p=0.42

# Power: could n=39 detect a true 60.5% rate? (two-sided, alpha=0.05)
n, p1 = len(df), 0.605
se0, se1 = np.sqrt(0.25/n), np.sqrt(p1*(1-p1)/n)
hi, lo = 0.5 + 1.96*se0, 0.5 - 1.96*se0
print((1 - stats.norm.cdf((hi-p1)/se1)) + stats.norm.cdf((lo-p1)/se1))   # 0.25

# Favorite: higher-ranked team (1994+, when rankings exist)
r = df.dropna(subset=["rank_first", "rank_second"])
higher = np.where(r["rank_first"] < r["rank_second"],   # lower number = higher rank
                  r["team_first"], r["team_second"])
print(one_prop_z((r["winner"].values == higher).sum(), len(r)))   # 15/31, p=0.86
```

A.2 Kick level: conversion, mechanism, and pressure

Each shootout is stored as its string of scored/missed bits in kick order; kicks alternate first, second, first, …, so the string plus "who kicked first" reconstructs every kick's team, result, and the running score. A kick is must-score if a miss would leave the team unable to catch the opponent, and to-win if scoring ends the shootout.

```
# per kick: side (first/second), scored (0/1), and pressure flags computed
# from the running score (full code in analysis_kicks.py)
def two_prop_z(m1, n1, m2, n2):
    p1, p2 = m1/n1, m2/n2
    pp = (m1 + m2) / (n1 + n2)
    z = (p1 - p2) / np.sqrt(pp*(1-pp)*(1/n1 + 1/n2))
    return p1, p2, z, 2 * (1 - stats.norm.cdf(abs(z)))

# mechanism: first- vs second-kicking team conversion
print(two_prop_z(125, 187, 122, 173))     # 0.668 vs 0.705, z=-0.75, p=0.45
# stakes: must-score kicks vs all others
print(two_prop_z(13, 26, 234, 334))       # 0.500 vs 0.701, z=-2.12, p=0.034
```

Appendix B: The Full Shootout Dataset

All 39 men's FIFA World Cup penalty shootouts, 1982–2026. "First kicker" is the team that took the opening kick; "Ranks" give the two teams' FIFA World Rankings at tournament time ("—" before rankings began in late 1992). The score is from the winner's perspective.

Table B1. All 39 men's FIFA World Cup penalty shootouts, 1982–2026.

Year	Round	First kicker → opponent	Winner	Score	Ranks (1st/opp)
1982	SF	France → West Germany	West Germany	5–4	—
1986	QF	Brazil → France	France	4–3	—
1986	QF	West Germany → Mexico	West Germany	4–1	—
1986	QF	Spain → Belgium	Belgium	5–4	—
1990	R16	Rep. of Ireland → Romania	Rep. of Ireland	5–4	—
1990	QF	Argentina → Yugoslavia	Argentina	3–2	—
1990	SF	Italy → Argentina	Argentina	4–3	—
1990	SF	England → West Germany	West Germany	4–3	—
1994	R16	Mexico → Bulgaria	Bulgaria	3–1	16/29
1994	QF	Sweden → Romania	Sweden	5–4	10/7
1994	Final	Italy → Brazil	Brazil	3–2	4/3
1998	R16	Argentina → England	Argentina	4–3	6/5
1998	QF	France → Italy	France	4–3	18/14
1998	SF	Brazil → Netherlands	Brazil	4–2	1/25
2002	R16	Rep. of Ireland → Spain	Spain	3–2	15/8

2002	QF	South Korea → Spain	South Korea	5–3	40/8
2006	R16	Ukraine → Switzerland	Ukraine	3–0	45/35
2006	QF	Germany → Argentina	Germany	4–2	19/9
2006	QF	Portugal → England	Portugal	3–1	7/10
2006	Final	Italy → France	Italy	5–3	13/8
2010	R16	Paraguay → Japan	Paraguay	5–3	31/45
2010	QF	Uruguay → Ghana	Uruguay	4–2	16/32
2014	R16	Brazil → Chile	Brazil	3–2	3/14
2014	R16	Costa Rica → Greece	Costa Rica	5–3	28/12
2014	QF	Costa Rica → Netherlands	Netherlands	4–3	28/15
2014	SF	Netherlands → Argentina	Argentina	4–2	15/5
2018	R16	Spain → Russia	Russia	4–3	10/70
2018	R16	Denmark → Croatia	Croatia	3–2	12/20
2018	R16	Colombia → England	England	4–3	16/12
2018	QF	Russia → Croatia	Croatia	4–3	70/20
2022	R16	Morocco → Spain	Morocco	3–0	22/7
2022	R16	Japan → Croatia	Croatia	3–1	24/12
2022	QF	Brazil → Croatia	Croatia	4–2	1/12
2022	QF	Netherlands → Argentina	Argentina	4–3	8/3
2022	Final	France → Argentina	Argentina	4–2	4/3
2026	R32	Germany → Paraguay	Paraguay	4–3	10/41
2026	R32	Netherlands → Morocco	Morocco	3–2	8/7
2026	R32	Australia → Egypt	Egypt	4–2	27/29
2026	R16	Colombia → Switzerland	Switzerland	4–3	13/19

The Imagined Gaze: Digital Beauty Standards, Celebrity Culture, and Eating Disorders Among Adolescent Girls By Martina Chai

Abstract

Taking the fatal case of a 15-year-old Chinese girl as a core entry point, this paper explores how traditional Chinese patriarchal culture, the K-pop industry's brutal thinness mandates, algorithm-driven social media content and commercial beauty industries collectively breed severe body anxiety and life-threatening eating disorders among adolescent girls. Centered on the concept of the imagined gaze, this research draws on social comparison theory and objectification theory to illustrate that teens with unstable self-identity easily internalize unattainable digitally curated beauty standards through parasocial idol bonds and viral extreme dieting trends such as the IU Diet and "What I Eat in a Day" videos, leading to constant upward comparison and self-surveillance that warp self-perception. Citing multiple psychological and public health studies, the paper outlines the irreversible physical and mental harms of anorexia, bulimia and orthorexia, and identifies shared complicity from neglectful families, unregulated social platforms, profit-driven diet brands and entertainment companies. I argue that media literacy alone cannot resolve the issue, and binding platform rules, algorithmic oversight and entertainment industry reforms are required to dismantle harmful thin beauty norms. Drawing on my own mild struggles with body image from identical digital pressures, I emphasize the case study's avoidable outcome as a stark reminder that countless young girls remain trapped in self-destructive cycles shaped by an externally constructed imagined gaze without large-scale structural intervention.

I. Introduction

There is a girl you may not have heard of. Her story was reported briefly in the Chinese press in spring of 2023, and then the news cycle moved on, as it always does. For the purposes of this paper, we will call her Ling.

Ling was fifteen years old, 165 centimeters tall, and in love with a boy in her class. She was the kind of girl who kept her feelings close, being observant, quiet, and alive in the way that adolescents usually are. And then, one ordinary day, she found out that the boy she liked preferred another girl. A girl who, amongst other things that Ling did not take into account, was thinner than her.

Having been raised by the Chinese media, Ling reached the conclusion that her body was the reason why he did not like her, and that his preference was a verdict on her worth. What Ling could not know—what perhaps none of us in that position can know—is whether his preference had anything to do with her body at all.

From cutting snacks, to refusing oil and staple foods, to refusing to eat altogether, Ling's caloric restriction intensified over months. In spring of 2023, she was wheeled into Shenzhen Children's Hospital after fifty days of consuming nothing but water. She was severely

malnourished, in respiratory failure, and unconscious. She died weighing 24.8 kilograms (China Newsweek).

I first encountered Ling's story on the news, and it stayed with me, partly because of its tragedy, and partly because of its familiarity. At the time, I was navigating my own, much less severe relationship with food and my body; one shaped by the same digital world, the same social pressures, and the same endless stream of images telling girls what they should look like. Ling's story began my search into why so many girls experience eating disorders, and what role social media and its digitally curated beauty standards play in that.

The question is not simple. Beauty standards do not fall from the sky or appear overnight. They are built by history, culture, media industries, algorithms, and the choices of millions of individual users to accept what they see as truth. Globalized media and celebrity worship have created a digitally saturated beauty standard prominent within Chinese youth culture that, by exploiting adolescent girls' developing sense of self and triggering chronic upward social comparison, produces irreversible harm—physical, psychological, and in some cases, fatal.

II. Ling's World

To understand what happened to Ling, we must first understand the world she grew up in. This is a world that has been telling girls for centuries that their value was inseparable from the smallness of their bodies. In China, the cultural association between femininity and thinness is not a modern invention. Visual and literary sources from imperial China have long normalized a thin ideal as representative of feminine attractiveness. A beauty print from the Qing dynasty explicitly lists "slim," "fair skin," and "a delicate body, and slender hands" as definitive traits of ideal female beauty, while *The Book of Songs* (诗经), the oldest surviving collection of Chinese poetry, identifies pale skin as the foundational element of a woman's physical beauty (Wang 64, 65). The most celebrated female character in classical Chinese literature, Lin Daiyu (林黛玉) from *Dream of the Red Chamber*, embodies this ideal to its extreme (Cao). Her beauty is inseparable from her sickly frailty, her pallor, and her emotional fragility. In the social context of the dynasties that produced these texts, to be beautiful was to be weak, and to be desirable was to be dependent.

Beyond aesthetic preferences, the ruling male-dominated class largely shaped China's beauty standard through Confucian ideology. As scholar Jaehee Jung notes, Confucian philosophy shaped traditional Chinese society by emphasizing women's "subordination and submissiveness to men," giving them little power or influence beyond the domestic sphere (Jung 59). In a culture where women were systematically taught that their worth lay in their appeal to men, physical delicacy became a form of currency. A woman's body was not her own; it was evidence of her social value. Before marriage, she was her father's property, and after marriage, she was her husband's property. Her appearance was proof of her worth to men, and her worth to men was proof of her existence. In this way, women were not only subjected to an external male gaze, they were taught to preemptively embody it—to see themselves as they imagined men would see them, so thoroughly that the distinction between the two became difficult to perceive.

This link between femininity, desirability, and physical fragility transcended classical texts and immortalized itself into the collective consciousness of Chinese women, persisting even today.

As Lei Zhang and colleagues document in their study *To Be Thin but Not Healthy*, the historical association of feminine beauty with fragility has evolved into a contemporary ideal among young Chinese women that prioritizes appearance over health, one in which "underweight has become synonymous with beauty" (Zhang et al. 2). Their study found that 76.15% of participants were pursuing an underweight BMI for aesthetic reasons rather than health, with one participant explicitly stating that her boyfriend wanted her to lose weight (Zhang 9). In a cultural situation where women were historically taught to value their appearance above their own well-being, adolescent girls have internalized the belief that their worth lies in their thinness, not their character, abilities, or actions. The obsession with slimness that shaped Ling's world was not an accident of her generation. It was centuries in the making.

III. The Screen in Her Pocket

History alone, however, does not explain what happened to Ling. Historical inheritance has always existed; what changed is the screen. In the digital age, beauty standards are no longer passed down inconspicuously through literature and family expectation. They are served to girls algorithmically, continuously, and at scale.

Ling grew up in a media environment saturated with images: K-pop idols presented as manifestations of physical perfection, body-checking trends that went viral overnight, and a thoroughly toxic social media ecosystem. As a Weibo poll of 56,140 netizens confirms, "the proliferation of weight loss methods on social media" was a primary driver of body image anxiety (China Newsweek). Jaehee Jung identifies China as a key market for Korean beauty industries due to the popularity of K-dramas, pop music, and entertainment (known as the "K-Wave" phenomenon). Korean youth culture has become an important influence on fashion and beauty trends, spreading through Chinese social media and broadcast platforms (Jung 58). The result is that K-pop idols, with their meticulously manufactured appearances and impossibly slender bodies, have become the de facto beauty standard for millions of young women across China and beyond.

Being thin is not simply a *feature* of K-pop idols, because they are products of an industry that enforces thinness with corporate precision. For example, TWICE's Momo was told to lose weight in a week if she wanted to appear in her group's debut showcase. BTS's Jimin survived on one meal every ten days before debut. IVE's Wonyoung, at the age of fourteen, was filmed consuming dieting products on the survival program Produce 48. Yuqi from (G)I-DLE shared in a variety show that she had once lost 4 kilograms in a single week by drinking only one cup of Americano per day, leading to severe hair loss. These cases are the documented[1] standard operating procedures of an industry that presents its performers as aspirational ideals while concealing the danger behind their production.

[1] One example of this is from the Chinese show 请吃饭的姐姐 (*Big Sister Who Invites You to Dinner*), in which Yuqi revealed her predebut dieting of 1 Americano per day.

Central to how these practices spread is the phenomenon of celebrity worship. Fans are susceptible to developing parasocial relationships with their idols. Through these one-sided emotional bonds formed through sustained media consumption, the celebrity's body becomes intimately interpersonal and comparable to their own. A study published in the *British Journal of Health Psychology* found that intense-personal celebrity worship was significantly associated with poor body image, specifically among female adolescents between the ages of 14 and 16, concluding that parasocial relationships with celebrities perceived as having a "good body shape" may lead to poor body image in young female fans (Maltby et al.). Additionally, a survey of 391 seventh- and eighth-grade students found that media exposure negatively predicted body image both directly and through a mediation process involving parasocial relationships with favorite characters and the motivation to self-compare (Young et al.). In other words, harm is caused not just by the exposure to thin images but also the emotional relationships girls form with the person embodying those images that makes the comparison feel personal and urgent. Consequently, the idol becomes a reference point for how the girl imagines others see her—not simply admired from a distance, but personally measured against it.

The most illustrative case of this social mechanism is the IU Diet. In 2013, Korean idol IU revealed her eating regimen on SBS's program Midnight TV Entertainment: one apple, two sweet potatoes, and a protein drink per day, totaling under 400 calories (SBS). To a person who has received education on health and nutritional facts, this diet may be immediately recognizable as unhealthy and detrimental to the human body. However, it rapidly went viral as the 'IU Diet Challenge,' enthusiastically adopted on YouTube primarily by young girls hoping to replicate her figure (South China Morning Post). A year later, IU disclosed on a talk show that she had developed bulimia, and had been binge eating and purging, caught in a cycle of anxiety and self-destruction (South China Morning Post). This reveal arrived too late to curb the trend, which had circulated at warp speed without any warning of its potential health impacts or how often a diet like this could be undertaken. The diet that launched thousands of imitations and became a collective aspiration had its origins in an eating disorder. Numerous YouTubers, predominantly young adolescent girls, tried the challenge publicly, filming themselves and their results, and documenting the headaches and weakness that followed (MUSE 247).

Imagine a girl—one composite of the many real stories that appear across social media comment sections, forum posts, and YouTube channels. As a devoted K-pop fan, she comes across the IU Diet. Like so many of her peers, she develops a parasocial relationship with an idol whose measurements are publicly discussed in fan forums, whose eating habits have been meticulously documented in press interviews, and whose body weight is routinely inspected and compared to the industry's beauty standard. The K-pop fan finds community in online spaces that not only reflect these aspirational images back at her, they also dictate what she imagines others see when they look at her. She begins to evaluate her own body as she believes her community would—to see herself not as she is, but as she fears she appears.

This is the environment that shapes girls like Ling. Mirroring the dynamics of imperial China, in which private shame was sustained through monitored femininity and family

expectations, contemporary beauty standards are packaged as lighthearted challenges, served by algorithms, and collectively celebrated in comment sections that read like communities, because for many of these girls, they are.

IV. The Media

If you are someone who is interested in watching dance, fashion, or beauty online, you may find, on any given day, videos titled ‘Daily Life of a 45kg Girl’ or ‘What I Eat in a Day.’ Clinicians suggest that the rise of WIEIAD videos across dance, fashion, and beauty spaces contribute to heightened eating disorder risk and harmful social comparison (Equip Health). In these videos, girls put their daily lives on display for their virtual friends to judge. Inspired by the girl who posted, viewers ask for links to her “dieting snacks.” They ask how she manages to keep her weight so low. They share their own goal weights. They call her beautiful. It is a girls-help-girls comment section full of viewers and contributors who recognize each other's beauty, in all the wrong ways. This is the social architecture of upward comparison, which Leon Festinger first theorized in 1954.

Festinger observed that “there exists, in the human organism, a drive to evaluate his opinions and his abilities”, an instinct to determine one's own worth through comparison with others (Festinger 117). In the digital age, the people we compare with are no longer peers or neighbors. They are curated, edited, professionally managed bodies, lit, filtered, and calibrated for maximum impact. The comparison is always upward, and always unwinnable. As Lei, an interviewee in Jaehee Jung's study, describes it:

“I think it's the media. They put those perfect female images in fashion magazines. Like, tell everyone this is the beauty, you should look like that. If you don't look like that, we won't like you. So everyone will feel like I'll never be skinny enough, I should be skinny.” (Jung 69)

What Lei is describing is the internalization of the idea that to be loved, you must be beautiful, and to be beautiful, you must be skinny. When the media perpetuates standards like these, girls like Lei and Ling become victims of a never-ending cycle of comparison and self-destruction.

Over time, this process of relentless upward comparison produces what objectification theory describes as self-surveillance: girls begin to internalize an external observer's perspective as their primary way of experiencing their own bodies (Zhang 11). They stop seeing themselves through their own eyes. But what replaces that self-perception is not simply the gaze of others—it is an imagined gaze: a constructed version of how they believe others see them, assembled from media images, social feedback, and the beauty standards they have spent years absorbing. This distinction matters. The external evaluation that a girl fears may bear little resemblance to how she is actually perceived; what shapes her behavior is the version of others’ judgment that she has built inside her own mind. What she imagines others see, filtered through everything the world has taught her about what a body should be, can become more real to her than the mirror. There are, in effect, three layers operating simultaneously: how she sees herself, how others around her actually see her, and how she imagines others see her. This last layer—the

imagined gaze—is often the most consequential of the three, precisely because it is immune to correction. It cannot be argued with. It can only be reinforced.

Research published in *Frontiers in Psychology* confirms that adolescents experiencing identity confusion—a developmentally normal lack of stable self-concept in early-to-mid adolescence—are significantly more susceptible to internalizing external beauty ideals and experiencing body dissatisfaction (Verschueren et al.). The less stable a girl's sense of self is, the more available she is to have it replaced by outside standards—and the more readily she fills the gap with imagined verdicts about how others perceive her.

This is precisely what was happening to Ling in the months before she stopped eating. She had received a piece of information—that the boy she liked preferred a different girl—and interpreted it through the logic her world had long been teaching her: that his preference meant she was not thin enough. Not being thin enough meant she was not beautiful or worthy of desire. Not being worthy of desire meant she was not a worthy person. She had absorbed that knowledge from centuries of cultural tradition and years of digital reinforcement, until her daily life seemed to confirm it. But we do not actually know how the boy she liked saw her. We do not know whether her body shaped his preference at all. What we know is how Ling imagined he saw her—a verdict assembled not from his actual perception, but from years of being told what kind of body earns the right to be chosen. The imagined gaze is not a reflection. It is a construction. And for Ling, as for so many girls, it was a more powerful reality than anything external.

A study published in *PLOS ONE* found that body-related social comparison was significantly more prevalent among adolescents with eating disorders than among their healthy peers, and that this association held even after controlling for depression and self-esteem, confirming that comparison itself—not merely the distress it produces—is part of the causal chain (Haedt-Matt et al.). This finding shifts the locus of responsibility away from the individual girl and her psychology, away from the framing of disordered eating as a personal failure of discipline or self-esteem, and toward the environments that normalize and amplify comparison: the comment sections, the challenge videos, and the parasocial idol culture. If comparison itself is pathogenic, then those environments are the conditions that produce eating disorders. Ling had developed her self-criticism inside a social media ecosystem specifically designed to generate exactly that kind of comparison.

V. The Body

Body dissatisfaction of the kind produced by chronic social comparison is consistently linked to anxiety, depression, and diminished self-esteem, particularly in adolescents, for whom these experiences occur during a critical developmental window. Research documents clear associations between media-induced body image disturbance and the onset of body dysmorphic disorder, a condition characterized by persistent, intrusive obsession with perceived physical flaws. The girls in those comment sections, asking for diet supplement links and saving thinspo

to private folders, are not engaging in harmless aspiration. They are rehearsing, at scale, the early cognitive patterns of disordered thinking.

The eating disorders that follow are medically serious and often overlooked by those experiencing the disorder and their families[2]. Anorexia nervosa involves severe restriction of food intake, rooted in distorted body perception and intense fear of weight gain. Bulimia nervosa involves cycles of bingeing and compensatory purging. Orthorexia is characterized by an obsessive fixation on 'clean' eating that crosses into dangerous restriction. Chen Jue, director of the Eating Disorders Treatment Center at Shanghai Mental Health Center, warns explicitly that among teenagers who diet to lose weight, 11.4% will develop binge eating within five years, and 3.1% will develop an eating disorder (China Newsweek).

The physical consequences extend well beyond thinness. Zhang et al. document that a BMI below 18.5 can reduce sex hormones and bone mineral density, leading to anemia, fatigue, low blood pressure, and cardiovascular risks (Zhang 2). Malnutrition in developing adolescent bodies does not only produce thinness, but also disrupts hormonal development, compromises organ function, and causes damage that may never fully reverse. Bones that do not mineralize properly in adolescence may not recover. Digestive systems that are disrupted by years of restriction or purging may not heal. Girls who survive eating disorders carry them forward in their bodies, and the disorder leaves a physical record.

VI. Everyone Is Complicit

What makes Ling's case particularly devastating is a detail from the *China Newsweek* report: her father did not initially think her restriction was serious. He “didn't think it would cause much harm and didn't pay much attention to it until she was diagnosed with anorexia nervosa” (Chinese Newsweek). He watched his daughter refuse oil and staple foods, and grow weaker, believing it was simply a “phase” that all girls go through at some point. The problem, however, is not a negligent father. The problem lies in the context he lives and is socialized in. The behavior of a girl slowly starving herself had been so thoroughly normalized that even someone who loved her, like her dad, did not recognize it as an emergency until she was dying. The beauty standard does not only work on the girls it targets. It works on everyone around them, teaching parents, teachers, and peers alike that restriction is discipline, thinness is virtue, and determination to be smaller is simply what girls do. The system requires this ambient acceptance to function. Without it, a primary pillar of cultural identification would collapse and subsequently, intervention would come sooner.

The media platforms are complicit. The Center for Countering Digital Hate documented that when 100 harmful eating disorder videos were flagged through YouTube's own reporting system, the platform failed to act on 83 of them, even while hosting advertising from global brands alongside that content (CCDH). Formulated from the combination of personal search behavior and trending topics, algorithms are designed to promote engagement, and from its

[2] This is demonstrated in the real life case of Ling, whose father expressed in an interview documented by *China Newsweek* that he did not think Ling’s eating habits were a big deal until she became hospitalized.

popularity in comments, saves, and shares, eating disorder content is highly engaging. This is particularly true in vulnerable communities, such as developing girls, who, as the algorithm has calculated, return to eating disorder content again and again. In a CBS News report one girl described how at twelve-years-old, she created a YouTube account to watch dog videos, and that over the following years, the platform's recommendations funneled her into eating disorder content. She now spends five hours a day consuming videos about body image. “YouTube became this community of people who are competitive with eating disorders,” she said. “And it kept me in the mindset that [anorexia] wasn't a problem because so many other people online were doing the same thing” (CBS News).

The beauty industry profits from the gap between the standard it sets and the body any individual girl inhabits. Clothing merchants deliberately make sizing charts smaller, so that girls, regardless of their weight, receive labels telling them they are large (China Newsweek). To take myself for an example, I weigh around 55kg, have a waist measurement of around 69-74cm depending on time of day, and am 171cm tall. In the US, I would wear a size S or M top[3]. According to a shop on Rednote, however, I would not even be able to fit in their XL size top, as the waist measurement for XL is 60cm[4]. These are how Chinese sizing charts play their part in the commercialization of young female insecurity, prepping teenagers to look for new products that can make them fit into these arbitrary sizes. Diet supplement companies target adolescents with products that promise quick and harmless results, yet they carry no scientific oversight. Brands sponsor influencers to post product content with buzzwords like ‘wellness’ and ‘natural’ that clinicians recognize as coded endorsement of disordered eating. Every element of this system is designed, at some level, to produce feelings of insufficiency among their primary audience, adolescent girls.

At its core, the problem is structural. It cannot be solved by telling girls to “love their bodies” while the systems that produce narratives of self-hatred remain unchanged. Media literacy is necessary, but insufficient without platform accountability. We cannot change the entire social media and entertainment industry quickly, nor can we fundamentally alter how generations of people have been taught to think about female beauty. What we can do is choose to reduce the exposure girls receive to content that promotes these dangerous ideals by establishing, reenforging, and requiring platform restrictions, algorithmic accountability, and advertising standards that treat eating disorder content as the public health hazard it demonstrably is.

Beauty standards are completely arbitrary. The standard shifts with whoever holds cultural, social, and economic power. The Qing dynasty celebrated frailty because a society that kept women subordinate needed to make that subordination look beautiful. The K-pop industry enforces thinness because a society that profits from girls' insecurity needs those insecurities to

[3] As according to JCrew’s sizing charts at https://www.jcrew.com/hk/r/size-charts?srsltid=AfmBOoqkj8oBYKNlH7MhWgcQ5IoQOfppN5CauCMhx-tZCGyvG-cwkPZb.

[4] From the store 糖五朵的店 (Tang Wuduo’s Store) on the social media app Rednote, the product is a simple spaghetti top and has over 44 thousand reviews with 0 ratings below a 4 star.

feel like personal failures rather than manufactured ones. Fashion brands, diet supplement sellers, and fast fashion retailers all capitalize on this manufactured inadequacy by marketing "wellness" and slim-fitting apparel to teens, and the normalization of this standard has made everyone—commercial players, families, and loved ones—complicit in sustaining toxic thin ideals. What is "beautiful" has never been a truth; it has always been a construction.

So what is the way out for an adolescent girl caught inside these systems?

Ling never got to find out. She never got to be sixteen, to get over her crush, to discover that his preference was not a verdict and her body was not a measure of her worth. She never got to exist in a body she made peace with, because the world she grew up in never gave her the tools to do that. That is a systemic failure, and without a change in our media, it is and will be the lived reality of millions of girls; girls who are, right now, in those comment sections, asking for the link to the supplement, saving the diet to a private folder, watching their own bodies through someone else's eyes. This is not a story about K-pop, or about China, or about social media in the abstract. It is a story about what we build, collectively, and who gets destroyed by it. I know, because it was partly my story, too. And because Ling's story should not have ended the way it did.

Works Cited

Cao, Xueqin, and E Gao. “Hong lou meng [Dream of the Red Chamber].” People’s Literature Publishing House, 1996.

Center for Countering Digital Hate (CCDH). “YouTube's Anorexia Algorithm: How YouTube Recommends Eating Disorder Videos to Young Girls.” counterhate.com, 10 Dec. 2024, counterhate.com/blog/youtube-pushes-harmful-eating-disorder-content-to-teens/.

China Newsweek. “15岁少女减肥至24.8公斤去世，什么加剧了身材焦虑” [“A 15-year-old girl died after dieting down to 24.8 kg. What exacerbated her body image anxiety?”]. Sina News, 25 May 2023, news.sina.cn/2023-05-25/detail-imyuyvkw5394754.d.html?vt=4. Accessed 14 Feb. 2026.

Equip Health. “The Hidden Harms of 'What I Eat in a Day' Videos.” Equip, reviewed by Jonathan Levine, LCSW, equip.health/articles/food-and-fitness/what-i-eat-in-a-day-videos-eating-disorder-risk. Accessed 23 July 2026.

Festinger, Leon. “A Theory of Social Comparison Processes.” Human Relations, vol. 7, no. 2, 1954, pp. 117–140.

Haedt-Matt, Alissa A., et al. “Body-Related Social Comparison and Disordered Eating among Adolescent Females with an Eating Disorder, Depressive Disorder, and Healthy Controls.” PLOS ONE, vol. 7, no. 10, 2012. PubMed Central, pmc.ncbi.nlm.nih.gov/articles/PMC3475236/.

Jackson, Todd, et al. “'Have You Taken the A4 Challenge?' Correlates and Impact of a Thin Ideal Expression from Chinese Social Media.” Frontiers in Psychology, vol. 12, 7 June 2021, doi.org/10.3389/fpsyg.2021.669014.

Jung, Jaehee. “Young Women's Perceptions of Traditional and Contemporary Female Beauty Ideals in China.” Family and Consumer Sciences Research Journal, vol. 47, no. 1, Sept. 2018, pp. 56–72, doi.org/10.1111/fcsr.12273.

Maltby, John, et al. “Intense-Personal Celebrity Worship and Body Image: Evidence of a Link among Female Adolescents.” British Journal of Health Psychology, vol. 10, no. 1, 2005, pp. 17–32, doi.org/10.1348/135910704X15257.

“Model Tries Crazy IU KPop Diet.” Project MUSE, Johns Hopkins University Press, muse.jhu.edu/pub/166/oa_edited_volume/chapter/3397909.

“SBS [Han-bam TV Entertainment] - It’s almost like fasting. IU’s diet menu is revealed!” YouTube, uploaded by SBS, 24 Oct. 2013, https://youtu.be/AjGagn1Wls8?si=s8pxk9FJ_PO1UZt0. Accessed 23 July 2026.

South China Morning Post. “K-pop Star IU's Battle with Anxiety and Eating Disorders, and How She Overcame Them.” South China Morning Post, 6 Jan. 2021, https://www.scmp.com/magazines/style/well-being/article/3116079/k-pop-star-ius-battle-anxiety-and-eating-disorders-and.

“YouTube Taught Me How to Have an Eating Disorder.” CBS News, 10 Dec. 2024, cbsnews.com/news/youtube-anorexia-algorithm-60-minutes/.

Verschueren, Michiel, et al. “Eating Disorder Symptomatology and Identity Formation in Adolescence: A Cross-Lagged Longitudinal Approach.” Frontiers in Psychology, vol. 9, 2018. PubMed Central, pmc.ncbi.nlm.nih.gov/articles/PMC5994691/.

Xiaoming Wang, Anita. “The Idealised Lives of Women: Visions of Beauty in Chinese Popular Prints of the Qing Dynasty.” Arts Asiatiques, vol. 73, no. 1, 2018, pp. 61–80, doi.org/10.3406/arasi.2018.1993.

Young, Andreanna F., Shira Gabriel, and Grainne Sechrist. “The Skinny on Celebrities: Parasocial Relationships Moderate the Effects of Thin Media Figures on Women's Body Image.” Social Psychological and Personality Science, vol. 4, no. 3, 2012, pp. 315–322.

Zhang, Lei, et al. “To Be Thin but Not Healthy — the Body-image Dilemma May Affect Health among Female University Students in China.” PLOS ONE, vol. 13, no. 10, 10 Oct. 2018, doi.org/10.1371/journal.pone.0205282.

Extending Lithium-Ion Battery Lifetime: A Review of Failure Prediction and Degradation Mechanisms By Arav Choudhary, Advisor By Sanjeev Datta

Abstract

Forecasting battery lifetime is a major challenge in battery science. Aging under real-world conditions is hard to evaluate as the data collection done in laboratories use ideal charging and discharging conditions which differ from dynamic conditions present in battery use. This review combines recent research in three areas related to building high quality batteries. The paper covers degradation processes present in Lithium-ion batteries (Guo et al. 2023) and how signs of failure can be obtained from early-cycle data measurements (Severson et al. 2019). Finally, it examines changes to standard industry testing protocols, such as dynamic discharge and pulse-current charging (Geslin et al. 2025; Guo et al. 2024) that can be used to improve the quality of a battery. These findings show clear needs: laboratory testing must align with real-world duty cycles to validate quality and that adaptive charging strategies being developed can be used to extend the lifetime and performance of lithium-ion batteries.

Introduction

Predicting battery cycle life focuses on identifying early metrics that reflect long-term battery degradation. Manufacturers aim to detect, quantify, and understand early indicators of future performance loss. To do this, they analyze resistance, voltage, and other features collected during the first few cycles of battery life. These early health metrics are then correlated with long-term degradation, enabling lifetime prediction early in production. Cells that are likely to fail can be identified before extensive field testing, improving reliability while reducing costly replacements. Beyond screening, these insights deepen our understanding of failure mechanisms enabling material, process, and design improvements. Ultimately, the goal is to deliver longer lasting batteries and a lower total cost of ownership for consumers.

This paper reviews recent advances in battery lifetime prediction. Machine learning (ML) models are being applied to large datasets of battery metrics. Studies show that cycle life can be predicted using only early-cycle voltage features from lab tests (Severson et al., 2019). These methods detect degradation signals much before traditional indicators, such as capacity loss, become visible. As a result, manufacturers can identify quality issues earlier, improve yield, accelerate validation, and enhance long-term reliability before the products reach the field.

Controlled battery tests often differ from real-world use cases. Electric vehicle (EV) batteries experience changing discharge rates due to driving behavior, traffic, and idle time, while lab tests typically use fixed charge-discharge cycles. Studies show realistic duty cycles can significantly affect degradation and battery life (Geslin et al., 2025). Battery degradation and aging occur due to multiple mechanisms, including interfacial film growth, electrode cracking, active material loss, and lithium inventory depletion (Guo et al., 2023). Understanding how operating conditions influence these processes is critical for improving reliability. Recent work shows that dynamic discharge profiles (Geslin et al., 2025) and pulse-current (PC) charging protocols (Guo et al., 2024) can significantly extend cycle life. This is challenging long-standing

assumptions in battery evaluation that steady state protocols are less damaging. The following sections examine these studies in detail. They explore how quality-control strategies and data-driven methods can identify degradation at an early stage. They describe the mechanisms behind battery failure and how these mechanisms interact with dynamic charge-discharge protocols. Together, these insights provide new pathways for extending battery life.

Part I: Data-Driven Early Prediction of Battery Cycle Life

This section details a data-driven approach from Severson et al. (2019) to predict battery life. The study successfully predicts the cycle life of commercial cells using electrochemical data from only the first 100 cycles, a point at which conventional degradation metrics had not yet appeared in battery health metrics. This application of combining early battery data with ML based methodology is a key innovation of the study. Such a data driven strategy is becoming increasingly useful for the variety of cell platforms and battery material systems in use today.

Experimental Design and the Limitations of Capacity-Based Prediction

An accurate predictive model requires a large and diverse dataset. To build one, the study cycled 124 commercial LFP/graphite cells under 72 charging protocols. This approach created a wide range of degradation behaviors, with cycle life spanning roughly 150 to 2,300 cycles. End of life was defined as the point where capacity fell to 80% of the nominal value. A key finding was that traditional battery health metrics are poor predictors of long-term performance. Figure 1 compares final cycle life with early indicators such as discharge capacity and capacity fade measured in the first 100 cycles. Only weak correlations were observed, showing that early capacity loss is not a reliable predictor of cell lifetime.

A Novel Feature: The Discharge Voltage Curve

The key innovation was moving beyond single measurements typical to battery tests such as capacity. Instead, the study analyzed the full discharge voltage curve. This was achieved by representing discharge capacity as a function of voltage, Q(V) and then calculating the difference between two distinct points in early-life: $\Delta Q_{100\text{-}10}(V) = Q_{100}(V) - Q_{10}(V)$. This transformation measures the per-voltage change in capacity between cycle 10 and cycle 100. This approach captured how the cell's electrochemical behavior evolved during early life. It revealed subtle degradation signals that traditional capacity metrics missed. A simple statistic of this new feature, its variance $\mathrm{Var}(\Delta Q_{100\text{–}10}(V))$, showed a remarkably strong relationship with final cycle life. Figure 2 shows a correlation coefficient of -0.92 between the log of this variance and the log of cycle life. Cells with larger voltage-profile changes in the first 100 cycles were destined for a shorter life.

Model Performance and Benchmarking

Researchers next built three machine-learning models to test predictive ability with different data sets. Model one used only the variance of $\Delta Q_{100\text{-}10}(V)$ as the critical feature for

prediction. Model two added additional candidate features obtained during discharge. Model three added data streams like temperature and internal resistance to the predictive set. A training set of 41 cells was used to select the model features and set the values of the coefficients for the features. A separate test set of 40 cells was then used to evaluate the models. Using data from only the first 5 cycles, a “full classifier” model reached 95.1% test accuracy. It correctly sorted cells into low- and high-lifetime groups. This contrasts sharply with capacity fade only models. For example, predicting with the training set's average cycle life gives about 36% average error on the test sets. The study shows that the model's value goes beyond quantitative prediction to support classification tasks, such as sorting cells into performance tiers on a production line.

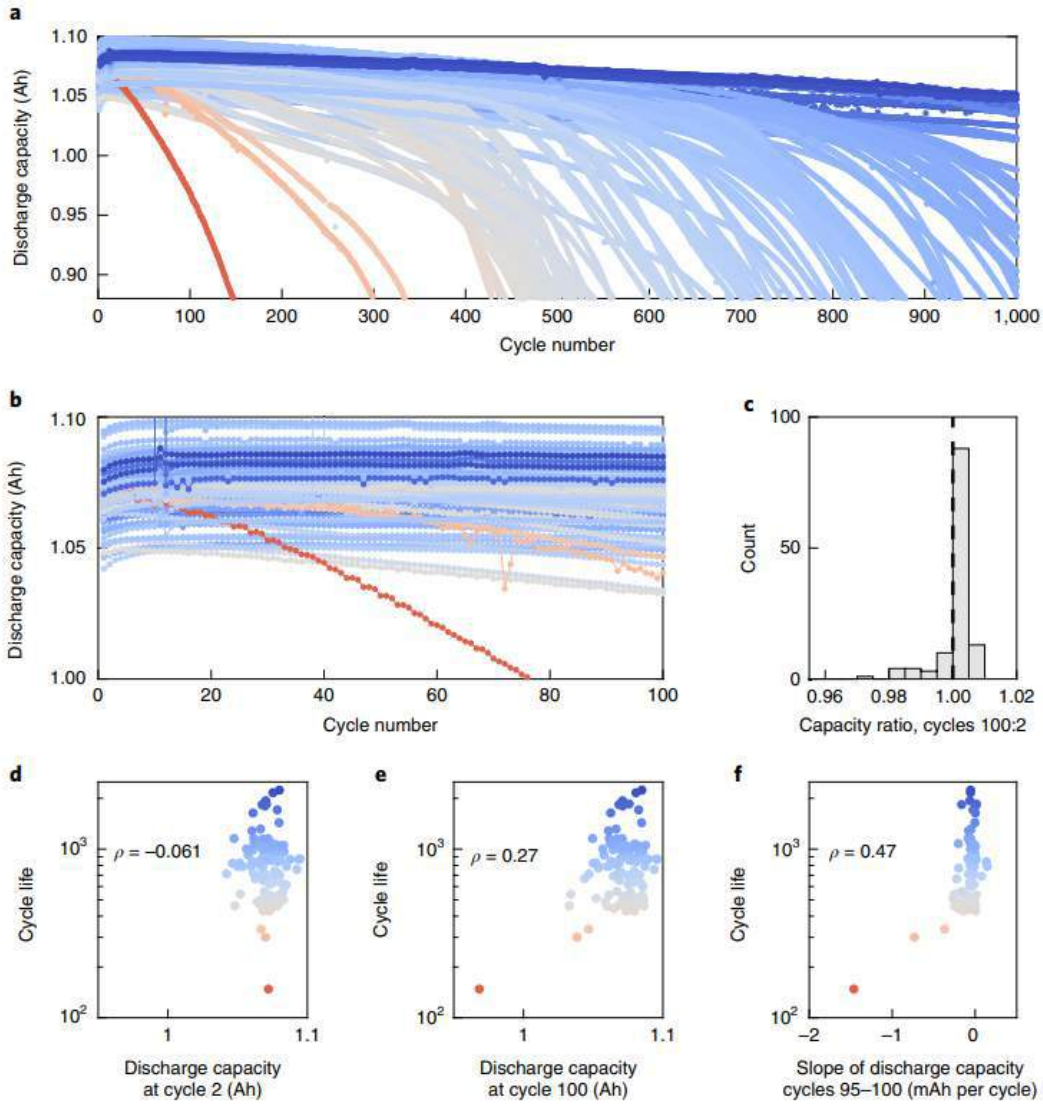


Figure 1: Poor predictive performance of battery end-of-life based on discharge capacity of the first 100 cycles. a) Discharge capacity for the first 1,000 cycles of LFP/graphite cells. b) A detailed view of a, showing only the first 100 cycles. c) Histogram of the ratio between discharge capacity at cycle 100 and cycle 2. d) Cycle life as a function of discharge capacity at cycle 2. e) Cycle life as a function of discharge capacity at cycle 100. f) Cycle life as a function of the slope of the discharge capacity curve .for cycles 95–100 (Severson et al., 2019)

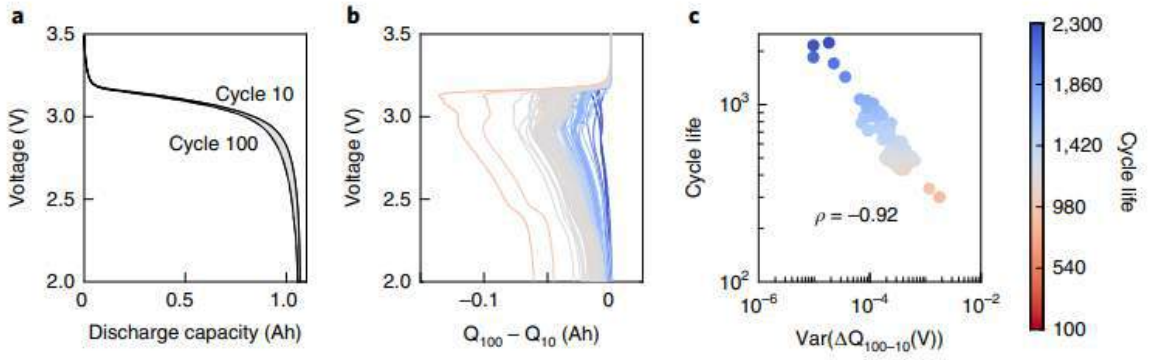


Figure 2: Early-Cycle Data-Driven Metrics. a) Discharge capacity curves for the 100th and 10th cycles for a representative cell. b) Discharge capacity difference curves as a function of voltage between the 100th and 10th cycles, $\Delta Q_{100\text{-}10}(V)$, for 124 cells. c) Cycle life as a function of the variance of $\Delta Q_{100\text{-}10}(V)$ with a correlation coefficient of −0.93 (Severson et al., 2019).

Rationalizing Predictive Success via Underlying Degradation Modes

The success of the voltage-based features is not just a statistical artifact. The authors provide justification based on physical degradation mechanisms occurring in batteries. Diagnostic analysis revealed that the ΔQ(V) feature is effective as it captures degradation modes that change the voltage curve long before they impact overall capacity. One such mode is Loss of Active Material of the Delithiated Negative Electrode (LAM_{deNE}), mapped by Dubarry et al. (2012). This behavior is observed when the negative electrode capacity is larger than that of the positive electrode, as is the case in these LFP/graphite cells. Here a loss of delithiated negative electrode material initially changes the potential without changing the overall capacity. This early, subtle change is what the ΔQ(V) feature detects. As LAM_{deNE} progresses, it eventually accelerates and capacity fades later in life. The model succeeds because it identifies the early symptoms of a degradation pathway, rather than waiting for the terminal symptom of capacity to fade.

Part II: The Influence of Dynamic Cycling on Battery Lifetime

Early cycle-life prediction is valuable for evaluating cells. But it is equally important to know if degradation seen in lab tests reflects real-world behavior. Most lab studies rely on constant-current cycling. This method differs greatly from the dynamic, variable loads batteries face in EVs. Real-world discharge rates depend on driving conditions, traffic, and usage patterns. This section explores a systematic investigation by Geslin et al. (2025) which challenges the constant-current testing convention used in lab protocols. It reveals that realistic dynamic cycling has a significant, and surprisingly positive, impact on battery lifetime.

Experimental Design: Simulating Real-World Operation

The Geslin et al. (2025) study was a large-scale, 24-month aging campaign. It used 92 commercial silicon oxide-graphite/nickel cobalt aluminum (SiOx-graphite/NCA) EV cells. The

cells cycled through 47 distinct discharge profiles. Discharge rates ranged from C/16 to C/2 designed to simulate real-world usage patterns. The profiles used in the study fall into four main categories. The first profile utilized constant current discharge prevalent in traditional laboratory baseline, with and without rest periods. The 2nd one was periodic and used simple dynamic profiles simulating segments of driving and regenerative braking. The 3rd one was synthetic, mimicking more complex profiles generated from real-world urban and highway driving data. The final profile used real full-day discharge profiles taken directly from the field.

Core Finding: Dynamic Cycling Enhances Battery Lifetime

Figure 3 compiles the aging test results across the various discharge protocols. The most impactful discovery is that constant-current profiles underestimate cell life compared to dynamic profiles for the same average discharge rate. Depending on the profile, dynamic discharge extended lifetime by as much as 38%. This finding contradicts a common assumption that fluctuating currents and high-power pulses from real-world driving are inherently more damaging to batteries. The results of this study suggest otherwise. Constant-current testing is not just unrepresentative; it is also overly conservative and can lead to the premature rejection of viable cell chemistry or to over-engineered battery packs.

The data further revealed an optimal discharge rate window for this cell chemistry between roughly 0.3C and 0.5C. Below this window (<0.3C), degradation is dominated by time-induced (calendar) aging, since cells simply spend more absolute time degrading. Above this window (>0.5C), degradation is dominated by cycling-induced stress. The existence of this optimal point has major implications for battery pack sizing and operating strategy design. The exact window likely depends on cell design, chemistry, and aging conditions, such as temperature or voltage window emphasizing the need for experiments that represent real field conditions. To identify the causes of lifetime differences, the researchers used ML on discharge-profile features. The strongest predictors of battery life were low-frequency current pulses (f_2), maximum discharge current, and the fraction of time spent resting at high state of charge (SOC).

Linking Data Metrics to Specific Degradation Modes

The study's final step linked measurable discharge-profile characteristics to physical battery degradation. Researchers extracted the capacities of the positive electrode (Q_{pe}), negative electrode (Q_{ne}), and lithium inventory (Q_{Li}) from periodic Reference Performance Tests (RPTs) run during the aging campaign. Correlating these capacities with the key dynamic features identified above showed that regardless of cycling protocol the battery degradation is initially dominated by the loss of lithium inventory (Q_{Li}). This mechanism corresponds to reactions involved with the growth of the solid electrolyte interphase (SEI) layer. The SEI layer is detailed later in this review. As batteries age, additional degradation mechanisms become important. The positive electrode capacity (Q_{pe}) degradation starts to dominate and was found to be directly linked to the rest fraction at high State of Charge (SOC). This aligns with known material science

principles (Gent et al. 2022), where the positive electrode is less stable and more prone to degradation at higher voltages (i.e., high SOC).

Q_{ne} degrades less than Q_{pe}. It is dictated by the maximum instantaneous discharge current and the current variance. Instantaneous current can reach up to 1,800% of the average discharge C-rate in real driving. The health of the silicon-containing negative electrode was strongly correlated with the effective depth of discharge (DoD). At low discharge C-rates, when the DoD is beyond 85%, the negative electrode capacity degrades more rapidly, while cells avoiding deep discharge have more preserved negative electrode capacities. Profiles with high instantaneous current peaks caused larger voltage drops, triggering the lower voltage cut-off earlier. This results in a shallower DoD, which protects the negative electrode by avoiding deep cycling into the low-SOC region where the silicon component is known to be electrochemically active (Kirkaldy et al. 2022). In addition, the DoD has no impact on the positive electrode capacity.

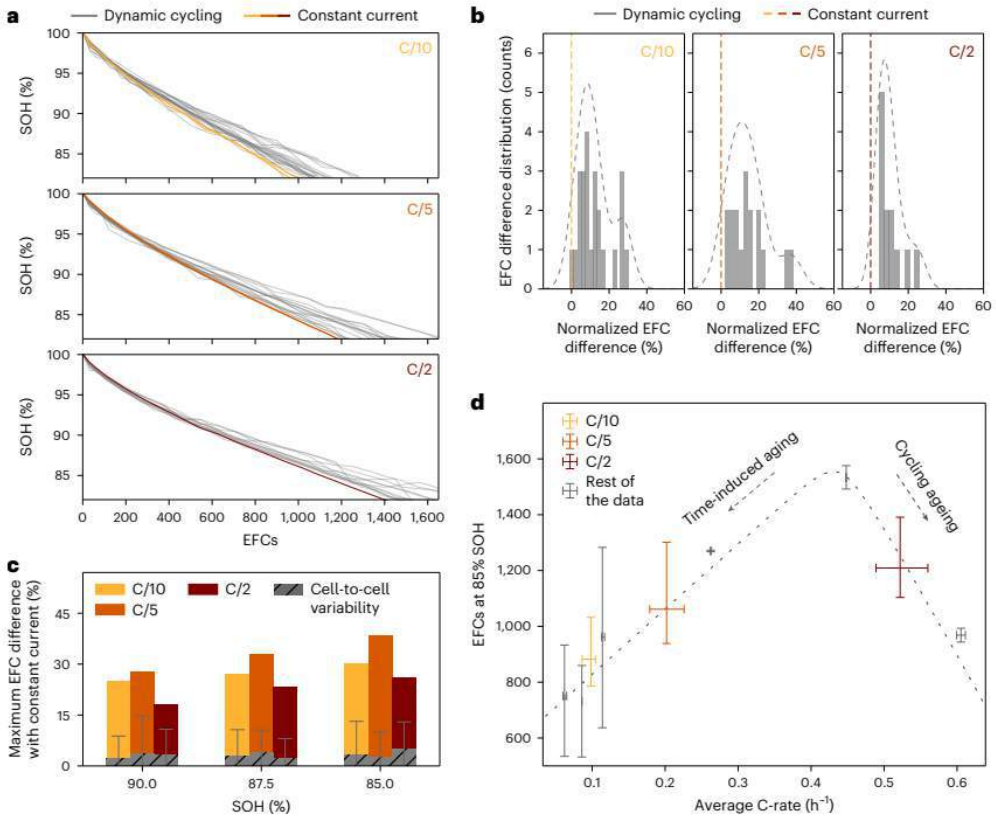


Figure 3: Impact of Dynamic Discharge Profiles on Battery Lifetime. a) Discharge capacity degradation trajectories, represented by State of Health (SOH), for cells cycled at C/10, C/5 and C/2. Constant current cycling profiles (in color) underestimate battery lifetime. b) Distribution of Equivalent Full Cycles (EFC) differences with respect to constant current cycling at 85% SOH for three different C-rates. c) Maximum EFC difference with respect to constant current cycling at 90%, 87.5% and 85% SOH for cells cycled at the same average C-rates. d) EFCs at 85% SOH as a function of the experimental average discharge C-rate (Geslin et al., 2025).

These findings reveal a nuanced picture. Different elements of a dynamic discharge profile can accelerate or mitigate specific aging mechanisms. This helps explain the wide variation in observed lifetimes. It also underscores the importance of testing batteries under conditions that mirror real-world use. This approach can define optimal operating windows within battery-management systems (BMS), tailored to each battery's specific material system.

Part III: Characterizing Degradation Under Constant Current (CC) Cycling

The previous section examined how discharge cycling rates affect cell health as measured electrochemically. This section builds on that foundation. It shows how measurements uncover key internal processes such as ion transport, charge-transfer kinetics, internal resistance, diffusion behavior, and electrode–electrolyte interfacial reactions which ultimately influence battery performance and aging.

Guo et al. (2023) ran battery-aging experiments with detailed post-mortem analyses. The goal was to identify the dominant degradation pathways in commercial $Li(Ni_{0.5}Co_{0.2}Mn_{0.3})O_2$ (NMC532)/graphite cells subjected to constant-current cycling. Although constant-current cycling is widely used, the mechanisms behind degradation are not fully understood. This study investigated the causes of capacity fade and resistance growth through loss of lithium inventory (LLI) and loss of active material (LAM). LAM in the NMC532 cathode and graphite anode was analyzed by making coin cells from aged electrodes. Resistance was measured using operando EIS in both full-cell and coin-cell configurations. LLI on the graphite anode was quantified using pH measurements and XPS analysis. The study linked LLI to interphase formation, its continued interphase growth, and lithium plating (Han et al., 2019). It also linked LAM to several known degradation mechanisms. These included transition-metal dissolution, graphite exfoliation, structural collapse, and the isolation of active material (Dubarry et al., 2012).

Electrochemical Performance and Polarization

Under CC cycling, the batteries showed a significant capacity drop. As the battery goes through repeated charging cycles, the RPT-measured charging plateau rises while the discharge plateau drops. This phenomenon is directly linked to the increase in internal impedance (Figure 4a). Incremental capacity analysis (ICA) quantifies this voltage plateau evolution (Figure 4b). In Figure 4b, peaks ‘a’ and ‘b’ correspond to phase transitions of active materials during charging and ‘c’ peaks correspond to transitions during discharging. The shift of ‘b’ peak over cycling to a lower potential is quantified to be the increase in polarization. It correlates with the large capacity fade observed under CC cycling expressed as state of health (SOH) in Figure 4c.

Impedance and Thermal Evolution

Operando EIS measurements of fresh and aged batteries appear as Nyquist plots in Figures 5a-c. These plots show three components representing the bulk resistance (R_{ohm}), SEI resistance (R_{sei}), and charge-transfer resistance (R_{ct}) (Zhu et al. 2020). R_{sei} relates to Li-ion diffusion at the interface while R_{ct} links to the electrochemical reaction. Both are known to

reduce with increased SOC (Wen et al. 2019). Bulk resistance of the battery is not expected to change with aging. EIS data show a four-fold increase in SEI resistance after 1000 cycles, compared to a fresh cell (Figure 5d). R_{ct} shows a threefold increase over the same period (Figure 5e). This indicates overall degradation at both the electrode interface and the charge-transfer with repeated cycling. The rise in resistance is accompanied with internal heat generation and surface temperatures reaching 55°C (Figure 5f). This heat further drives parasitic side reactions and may also rupture the interfacial film at high temperatures. The operando EIS data show that capacity fade during operation is dominated by rising internal resistance.

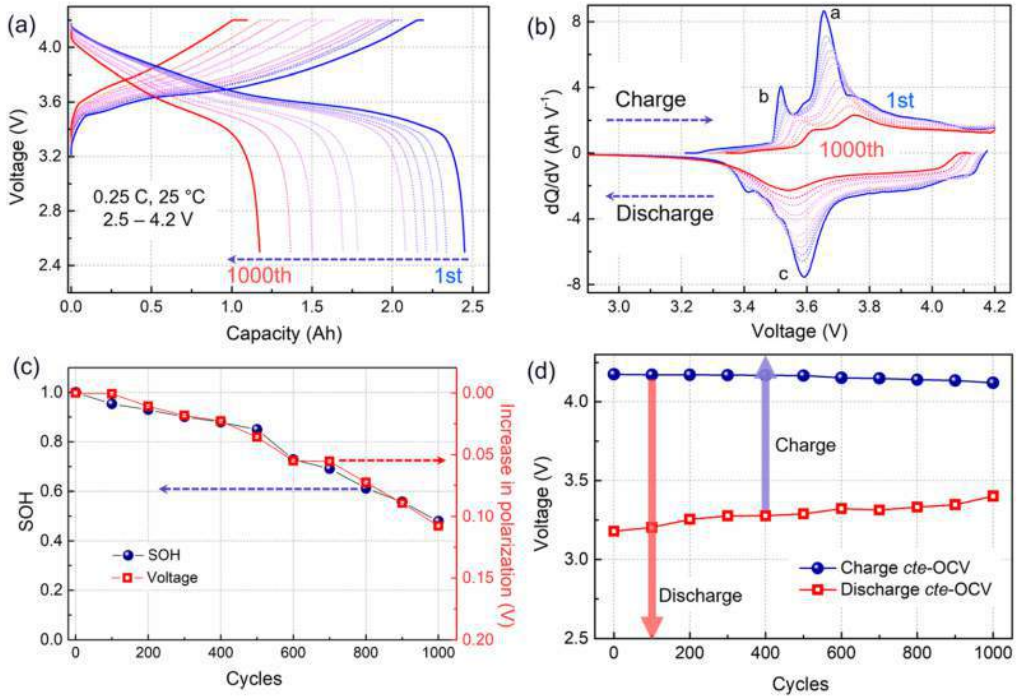


Figure 4**:** Electrochemical diagnostic results of cells under the CC protocol. Electrochemical behavior from RPT after every 100 cycles at 0.25C. a) Voltage profiles showing plateau shifts. b) ICA curves showing peak shifts due to polarization. c) SOH and polarization evolution. d) Close-to-equilibrium OCV trends (Guo et al., 2023).

Quantification of Lithium and Material Loss

Researchers disassembled fresh and aged cells to evaluate lithium loss, active material loss, and impedance increase. Lithium loss on the negative electrode, through Li plating or on the surface film, reduces the capacity of the battery. SEM images (Figures 6a-c) show the transformation of smooth graphite particles of the fresh negative electrode to a collapsed structure with a rough surface layer on the aged electrode. Figures 6d-g detail a novel pH-measurement and XPS-analysis method used to characterize the Loss of Lithium Inventory (LLI). This irreversible loss, from SEI formation and Li plating on the graphite anode, accounted for about 16.8% of total capacity loss for the battery. Loss of Active Material (LAM) was also evident at both electrodes. Post-mortem analysis of the NMC532 cathode revealed particle pulverization, grain volume increase, and cracking. It also revealed transition metal (Ni, Co, Mn)

dissolution which then deposits on the anode, accelerating its exfoliation (Figures 7a-h). This LAM causes a 14% capacity loss on the positive electrode (Figure 7i). The negative electrode also thickens with cycling, by 30.5%, and develops many cracks and exfoliated pieces (Figures 8a-f). Raman measurements of aged graphite show a loss of peaks (Figures 8g-h) indicating structural collapse with cycling. The stripped materials leave the current collector leading to R_{ct} increase (Figure 5e). Coin cells made from aged graphite electrodes show significant capacity loss during de-lithiation, compared to a fresh electrode confirming the stripping (Figure 8i).

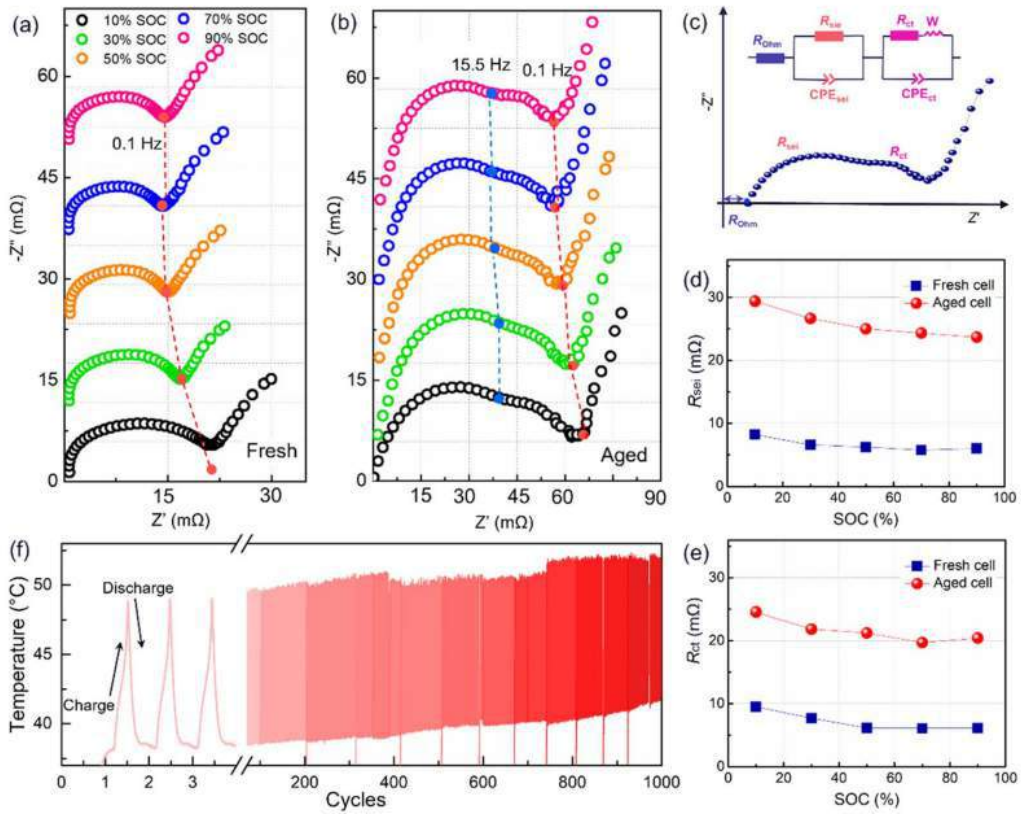


Figure 5: EIS Nyquist plots for fresh vs. aged cells across various SOCs, alongside real-time temperature fluctuations. EIS spectra at various states of charge for a-b) fresh vs. aged cell Nyquist plots; c) schematic illustration and an equivalent circuit used to fit the EIS Nyquist plots; fitted results for d) R_{sei} and e) R_{ct}; f) temperature measured with a sensing probe on the outer surface of the cell (Guo et al., 2023).

Interfacial Impedance of Electrodes

Li compounds plate onto the negative electrode surface. They combine with other solvents to form a SEI layer. Similarly, surface reactions on the positive electrode form a "cathode-electrolyte interphase" (CEI). The SEI and CEI layers create a thermodynamic potential gap between electrodes and electrolyte which results in an impedance increase, and thus capacity fade, with cycling. Figures 9a-c quantify this impedance change for the cells in the study. XPS analysis of the SEI and CEI layers reveals the compositional changes of the electrodes with cycling (Figures 9d-f). The SEI film gains more organic compounds with aging,

such as C-O and CO, which originate from electrolyte decomposition. These additional species at the interface reduce the percentage of Li+ available for diffusion leading to a loss of capacity. The CEI film contains $Li_xP_yOF_z$ and other similar compounds, which increase with aging. This comes from the electrolyte $LiPF_6$ decomposition at high voltage on the positive electrode.

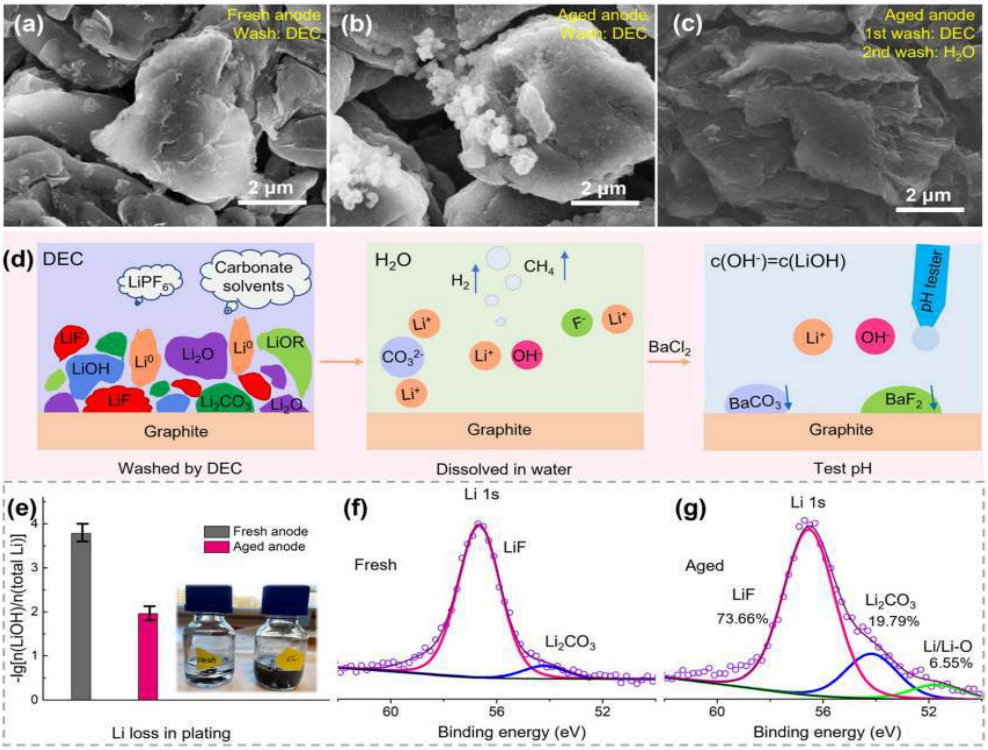


Figure 6: Methodology for quantifying inactive lithium on the anode via water reaction and pH testing. SEM images of the graphite anode from a) the fresh cell, b) the aged cell after washing with DEC, and c) the aged cell after washing with DEC and then with water; d) mechanism diagram of the dissolution of lithium compounds in water, quantified by the pH test, with the quantified LiOH content shown in e); Li 1s XPS spectra of the negative electrode from f) the fresh cell and g) the aged cell (Guo et al., 2023).

Perspectives for Increasing Stability of Battery Materials and Cell Life

Guo et al. (2023) show that under CC charging, impedance growth from electrode degradation is the major driver of battery life reduction. To extend battery life, the study recommends several options spanning full-cell operation, electrode fabrication, and electrode material preparation.

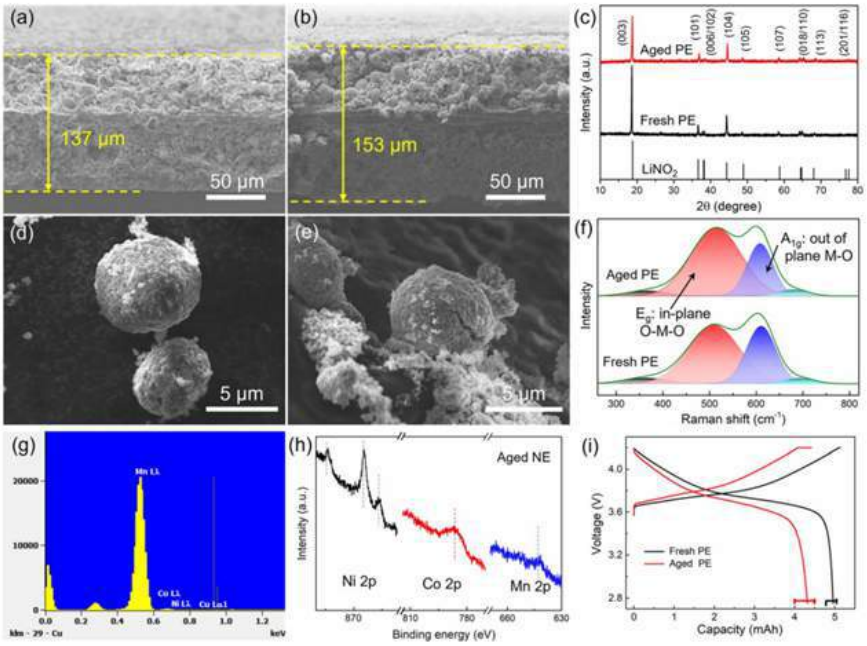


Figure 7: Structural cracking of the positive electrode. SEM images of a) fresh b) aged cathode. c) XRD of fresh and aged cathode. SEM images of d) fresh e) aged NMC particles. f) Raman spectra of the fresh and aged cathode. Transition metals (Mn, Co, and Ni) from positive electrode deposit on negative electrode, measured by g) EDS point scan and h) XPS. i) Voltage profiles of the fresh and aged positive electrodes (Guo et al., 2023).

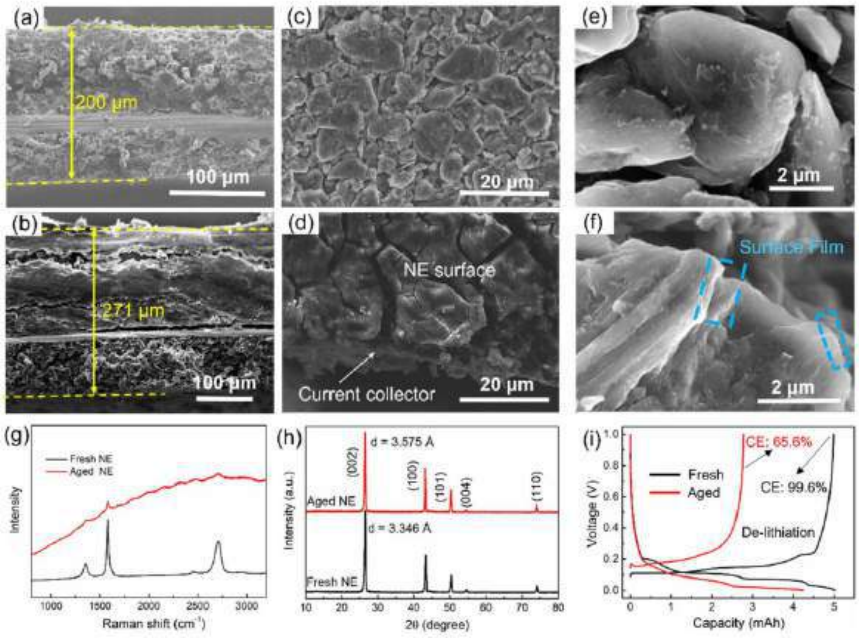


Figure 8**:** Structural degradation of the negative electrode showing anode exfoliation and thickening. Cross-section SEM images of a) fresh and b) aged negative electrode; top-surface SEM images of c) fresh and d) aged negative electrode; SEM images of e) fresh and f) aged graphite particles; g) Raman spectra of fresh and aged negative electrodes; h) XRD patterns of aged and fresh negative electrodes; i) charge and discharge voltage curves tested against Li in coin cells from 0.001–1.0 V at 0.05C (Guo et al., 2023).

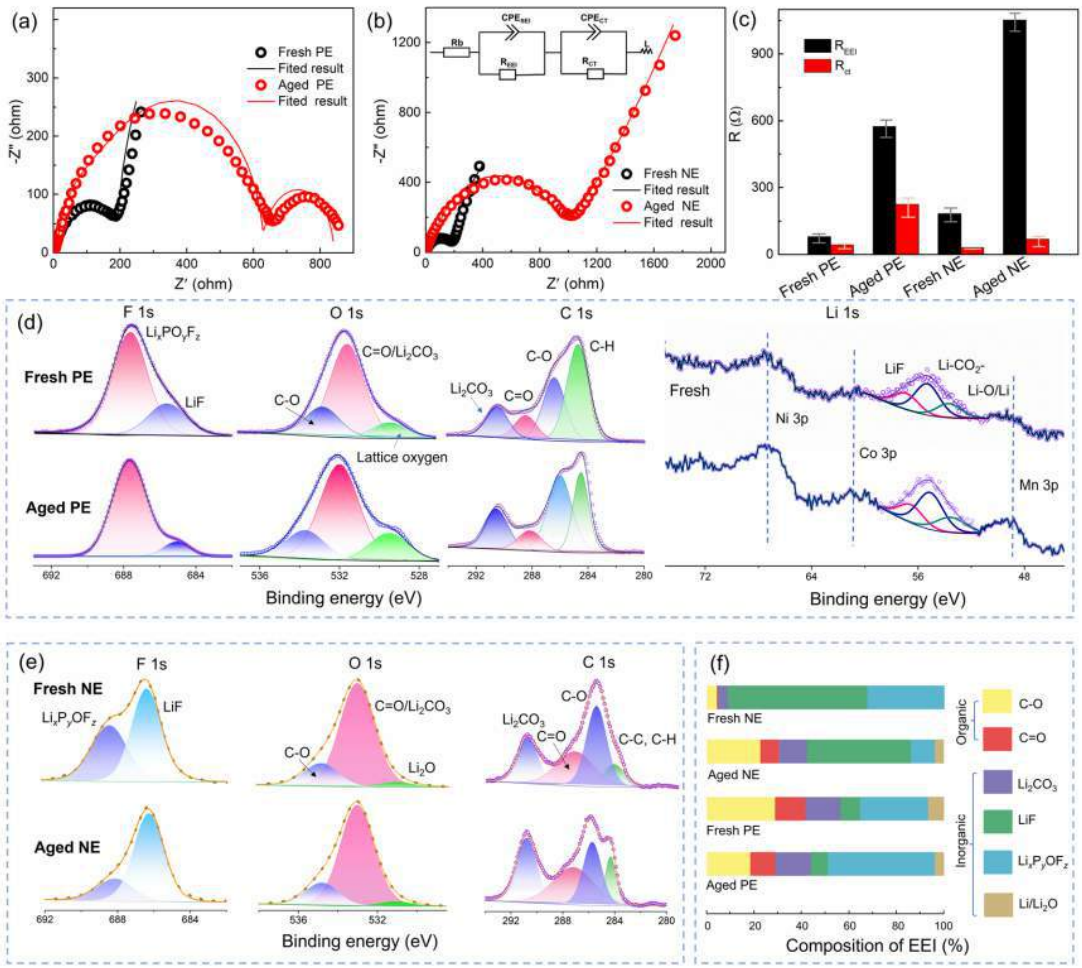


Figure 9: Quantitative XPS and impedance analysis of the SEI and CEI layers. Impedance data for a) positive electrode (PE) and b) negative electrode (NE), and c) fitted results, R_{EEI}: R_{sei} or R_{cei}; XPS spectra of d) F 1s, O 1s, C 1s, and Li 1s for the positive electrode and e) F 1s, O 1s, and C 1s for the negative electrode; f) chemical composition from quantitative XPS characterization of the SEI film on the negative electrode and the CEI film on the positive electrode (Guo et al., 2023).

Dynamic pulse current (PC) charging (Huang et al. 2021), instead of CC charging, can inhibit SEI thickening from repeated formation and cracking and increase battery life. Heat dissipation or cooling systems can reduce electrode temperature and slow down reaction rates. At the electrode level, a suitable binder can reduce peeling by forming a tight connection between active materials and the current collector. The ratio of negative to positive electrode capacity can also be optimized to balance SEI formation and Li-plating (Han et al. 2019). At the material level, single-crystal cathodes can lower the interface surface area. This reduces transition metal dissolution into the lattice, enhancing life (Guo et al. 2022). For the interface film, an artificial SEI on the graphite surface can inhibit material deposition, creating a stable electrode interface (Li et al. 2018). While the degradation mechanisms identified in this study and their mitigation

strategies may be specific to commercial NMC/graphite batteries the diagnostic methods can be applied to other battery chemistries supporting broader advances in durability and performance.

Part IV: Characterizing Lifetime Improvements Under Pulse Current Cycling Protocol

Parts I-III examined degradation metrics, how to measure and predict them across real-world charge-discharge rates, and the mechanistic foundations of failure. This section advances the discussion further. It explores whether charging strategies beyond conventional constant-current methods can extend battery lifetime. Various studies have reported SoH improvements of 0.5% to 34% using PC protocols across different commercial battery formats (Monem et al. 2015; Kannan et al. 2021; Li et al. 2019). However, the mechanisms behind these improvements have historically been unclear, hindering further optimization. Guo et al. (2024) studied the electrochemical behavior of 18650-type NMC532/graphite batteries by aging cells under two protocols: CC and PC. Detailed electrochemical analysis, post-mortem analysis, and operando characterization were performed to understand how aging processes change under each protocol

Battery Performance and Electrochemical Diagnostics

Figure 10a shows results for NMC532/graphite batteries under CC and two PC protocols: 100 Hz (Pulse-100) and 2000 Hz (Pulse-2000). After 1000 cycles, the CC protocol retained only 37.8% of initial capacity. Pulse-2000 retained 81.73%, more than double the CC result while Pulse-100 retained 66.5%. To understand the electrochemical markers behind this change, the study quantified battery polarization. The extent of polarization was found to decrease with pulsed current charging and decreased further as the pulse frequency increases. Figures 10b-d show this as a smaller shift in the secondary peak (Peak II) of the incremental capacity (IC) curve, a known way to quantify polarization (Dubarry et al. 2006). The main peak (Peak I) indicates loss of active materials in the cathode and anode electrodes (Guo et al. 2022) and also shows a significant reduction for the PC protocols. EIS and Distribution of Relaxation Times (DRT) analysis (Iurilli et al. 2021) show that PC charging produces lower interface (R_{sei}) and charge-transfer (R_{ct}) resistances compared to CC charging (Figure 11). Together, the IC, EIS, and DRT measurements show that PC charging mitigates SEI growth, reduces charge-transfer resistance, and eases the loss of electrode materials.

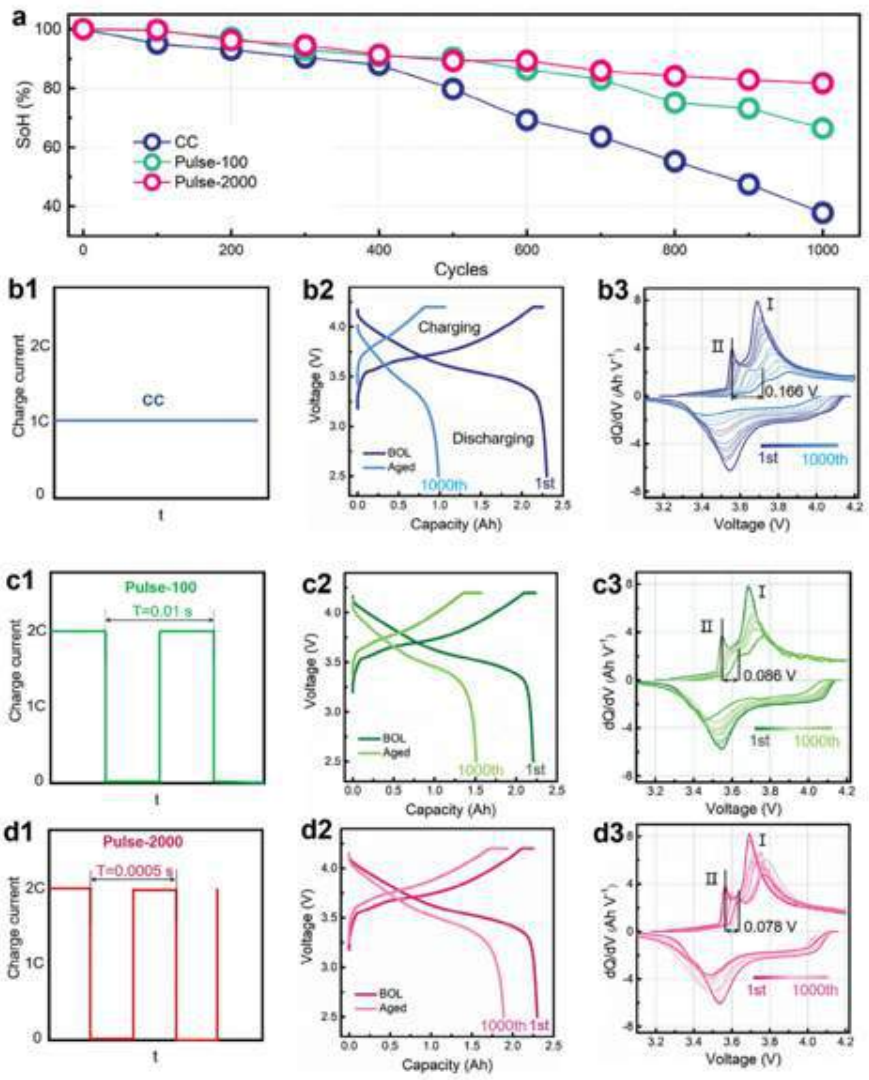


Figure 10**:** Electrochemical behaviors during RPTs for batteries aged with CC, Pulse-100, and Pulse-2000. RPT rests were performed after every 100 cycles of aging. a) Battery SoHs from the RPTs during aging. Current waveforms for battery aging, and dis-/charge voltage profiles and IC curves of the RPTs for batteries aged with b1-b3) CC, c1-c3) Pulse-100, and d1-d3) Pulse-2000 charging protocols. BOL: beginning of life (Guo et al., 2024).

Mechanism of Electrode Degradation Reduction

Post-mortem analysis was done by the researchers to understand the chemical and physical processes behind the dramatic Pulse-2000 lifetime gain over standard CC charging. All ex-situ measurements in the study confirmed that electrodes were more stable under PC charging. Synchrotron XRD revealed that PC charging reduces the cracking and pulverization of the graphite anode. This appears as a slower growth rate in electrode thickness for PC charged batteries (Figures 12a-b). Graphite undergoes complex phase transitions during Li-ion intercalation, such as LiC_{48}, LiC_{24}, LiC_{18}, LiC_{12}, and LiC_6, to reach a lower-energy state (Asenbauer et al. 2021). Operando Raman spectroscopy captured these intermediate phases, labeled I through IV (Figures 12c-d) which show that lithium-ion distribution is more uniform

under PC lithiation. This uniformity reduces mechanical stress and volume fluctuations in the electrode.

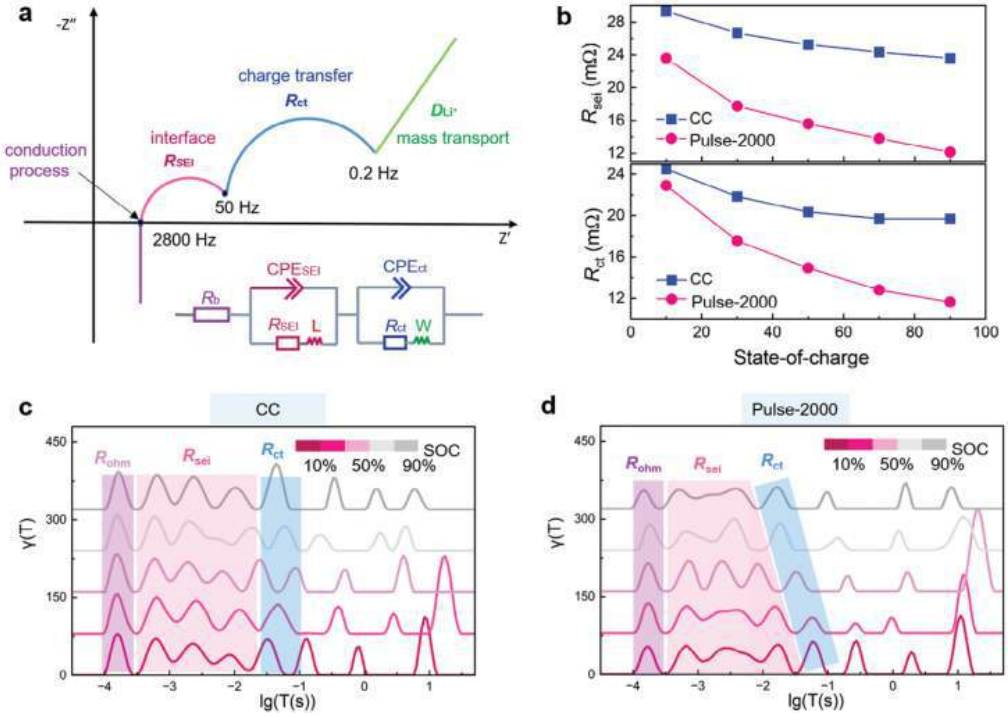


Figure 11: EIS and DRT comparison of batteries under CC and PC charging protocols. EIS results after 1000 cycles under CC and Pulse-2000 charging. a) Characteristic Nyquist plot and equivalent circuit diagram. b) Fitting results of R_{sei} and R_{ct} relative to SoC. Distribution of relaxation time (DRT) profiles at 10%, 30%, 50%, 70%, and 90% SoC for batteries aged under c) CC and d) Pulse-2000 charging (Guo et al., 2024).

The CC-aged cathode showed a greater change in crystal structure than the Pulse-2000 cathode. Ex situ synchrotron XRD, known to capture crystal volume expansion and layer spacing changes in NMC532 materials (Ates et al. 2014), showed a larger change for CC-aged cathodes than for PC-aged ones (Figure 13a). Extended X-ray Absorption Fine Structure (EXAFS) determines local atomic structure and coordination number in crystals (Yu et al. 2021; Ruess et al. 2022). EXAFS showed a larger change for the CC-aged cathode than the PC-aged cathode (Figure 13b). Operando X-ray absorption spectroscopy (XAS) and X-ray absorption near-edge structure (XANES) show that PC charging helps the Ni element reach a higher energy state (Figures 13c-d), with less Ni-O bond length variation, preserving the layered structure during the redox reaction (Lee et al. 2022).

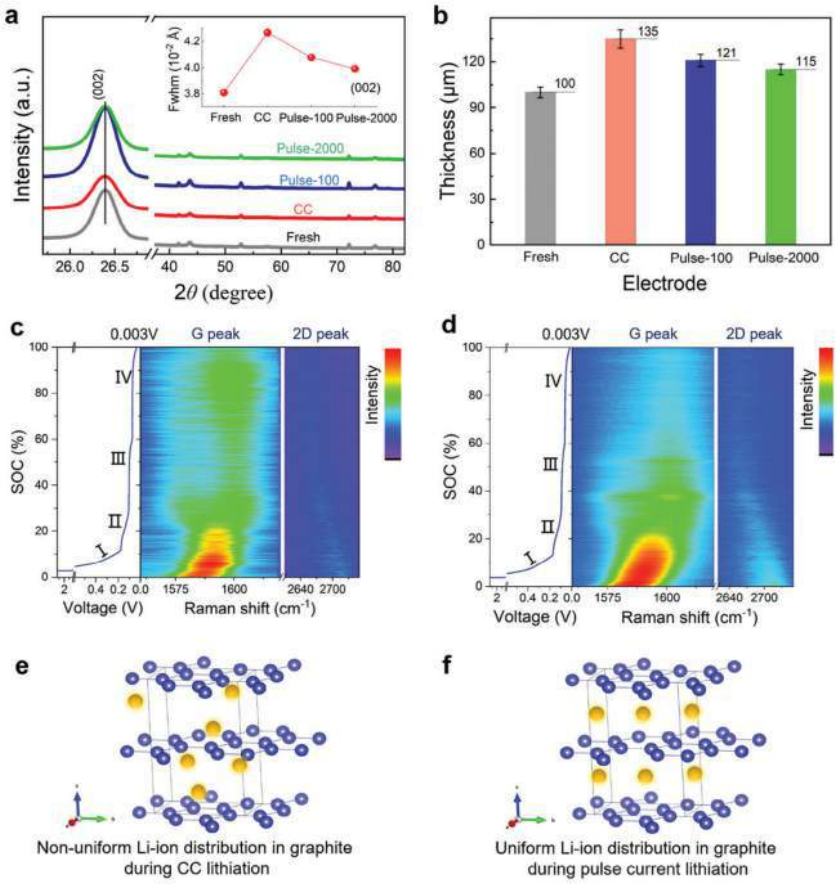


Figure 12: Graphite anode degradation: PC vs. CC protocols. a) Ex situ synchrotron XRD patterns of fresh and aged graphite anode, with the FWHM of the (002) peak inset; b) thickness of fresh and aged negative electrodes from SEM results; operando Raman spectra of graphite during c) CC and d) pulse current lithiation; state of graphite with Li-ions e) unevenly distributed during CC lithiation and f) evenly distributed during pulse current lithiation (Guo et al., 2024).

Graphite-Electrolyte Interface Growth Reduction

The authors next provide an explanation why PC charging suppresses SEI growth better than CC charging. Figures 14a-b compare Raman spectra of Pulse-2000-aged and CC-aged graphite anodes. The PC-aged anode shows sharper peaks, indicating better retained crystallinity. TEM and SEM images (Figures 14c-f) reveal more than a 2x reduction in SEI thickness growth for Pulse-2000 versus CC (50 nm vs. 110 nm). SEI formation and growth involve complex reactions of the decomposition of the electrolyte solvent and Li salt (Heiskanen et al. 2019) shown below.

$$EC + e^- \rightarrow LEDC + CH_2CH_2 \quad (1)$$

$$LiPF_6 \rightarrow Li_2CO_3 + Li_xF_yO_z \quad (2)$$

$$LEDC \rightarrow Li_2CO_3 + Li_2O + LiOR \quad (3)$$

XPS depth profiling (Figure 15a) showed that the SEI formed under PC charging is organic-enriched and contains fewer brittle inorganic salts, such as Li_2CO_3, LiF, and Li_2O. The authors propose that PC charging inhibits the reduction reaction between solvent and interface electrons (Figure 15b). This slows the initial reactions (Equations 1 and 2) as well as the secondary decomposition reaction (Equation 3). This consequently results in slower accumulation of inorganic species in the SEI, and a thinner SEI film overall.

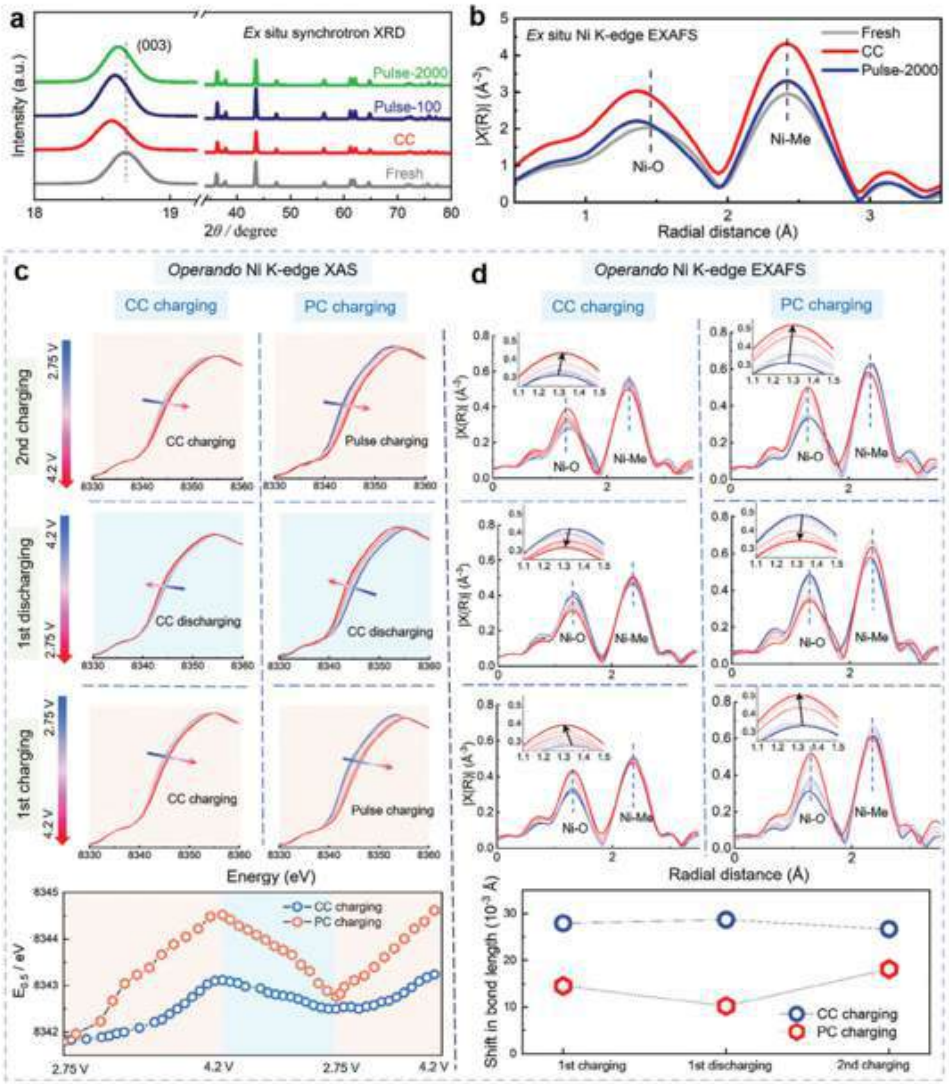


Figure 13: NMC cathode degradation: PC vs. CC protocols. a) Ex situ synchrotron XRD results for fresh and aged NMC cathodes, and b) Ni K-edge ex situ EXAFS; operando XAS for the fresh positive electrode under CC and pulse current charging; c) Ni K-edge XANES evolution of NMC532 during charging and discharging with CC and pulsed current; d) operando Ni K-edge EXAFS and Ni-O radial distance under CC and pulse current charging (Guo et al., 2024).

This study provides a detailed explanation of how PC protocols extend battery cell life. Electrochemical analysis, operando and ex situ characterization at the full-cell, electrode, and material level all provide mechanistic proof that high-frequency (2000 Hz) pulse current charging stabilizes NMC532/graphite batteries. PC charging is shown to balance lithium-ion concentration gradients at the interface and allows structural relaxation of the electrode. This minimizes mechanical strain, suppresses electrolyte decomposition, and preserves electrode crystallinity. Similar efforts are being undertaken using pulsed-current charging in other battery chemistries. One recent study applied pulsed-current formation protocols during metallic lithium electroplating in an anode-free Li-metal coin cell. This directly enhanced Li-metal stability in architectures relevant to solid-state batteries (Cicvaric et al. 2024). An ongoing research program at Aarhus University is investigating how PC waveforms affect Na-ion intercalation kinetics and cycling stability, with results expected later this year. This work will address how the larger Na+ ionic radius with different diffusion characteristics get impacted by pulse-charging. This study will be critical in understanding if the lithium-ion insights transfer directly to a system distinct from the SEI chemistry studied in this review (Guo et al. 2026).

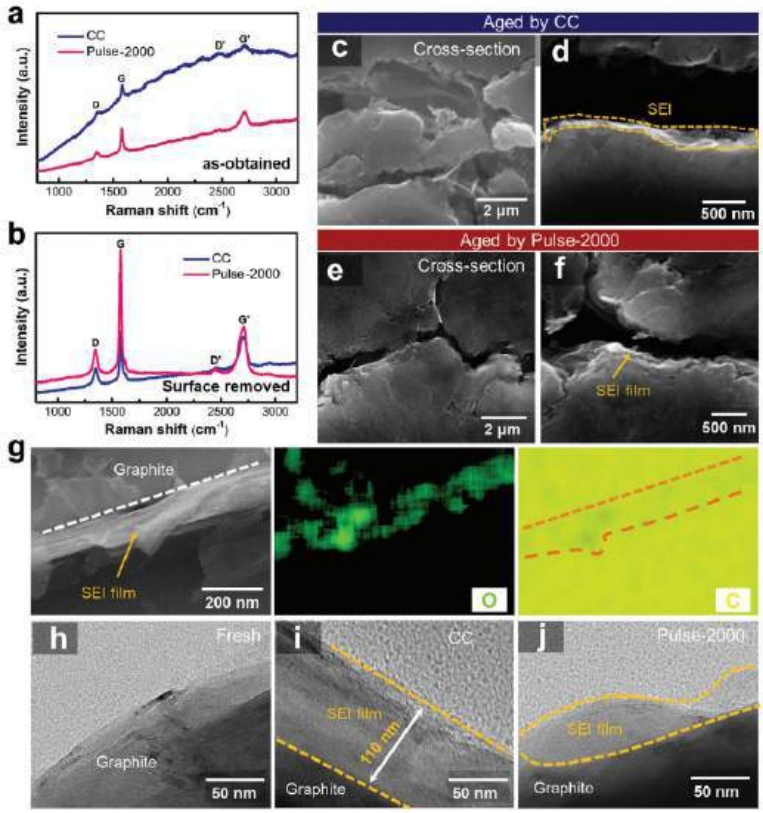


Figure 14: Graphite-electrolyte interface analysis. Raman spectra of a) as-cycled and b) surface-removed graphite anodes aged under CC and Pulse-2000 charging. FE-SEM images of cross sections of graphite electrodes aged with CC (c,d) and Pulse-2000 (e,f) charging; d,f) are edge-magnified images of c,e). g) Micrograph and O and C element mapping of the CC-aged graphite electrode surface. TEM images of h) fresh, i) CC-aged, and j) Pulse-2000-aged graphite anodes (Guo et al., 2024).

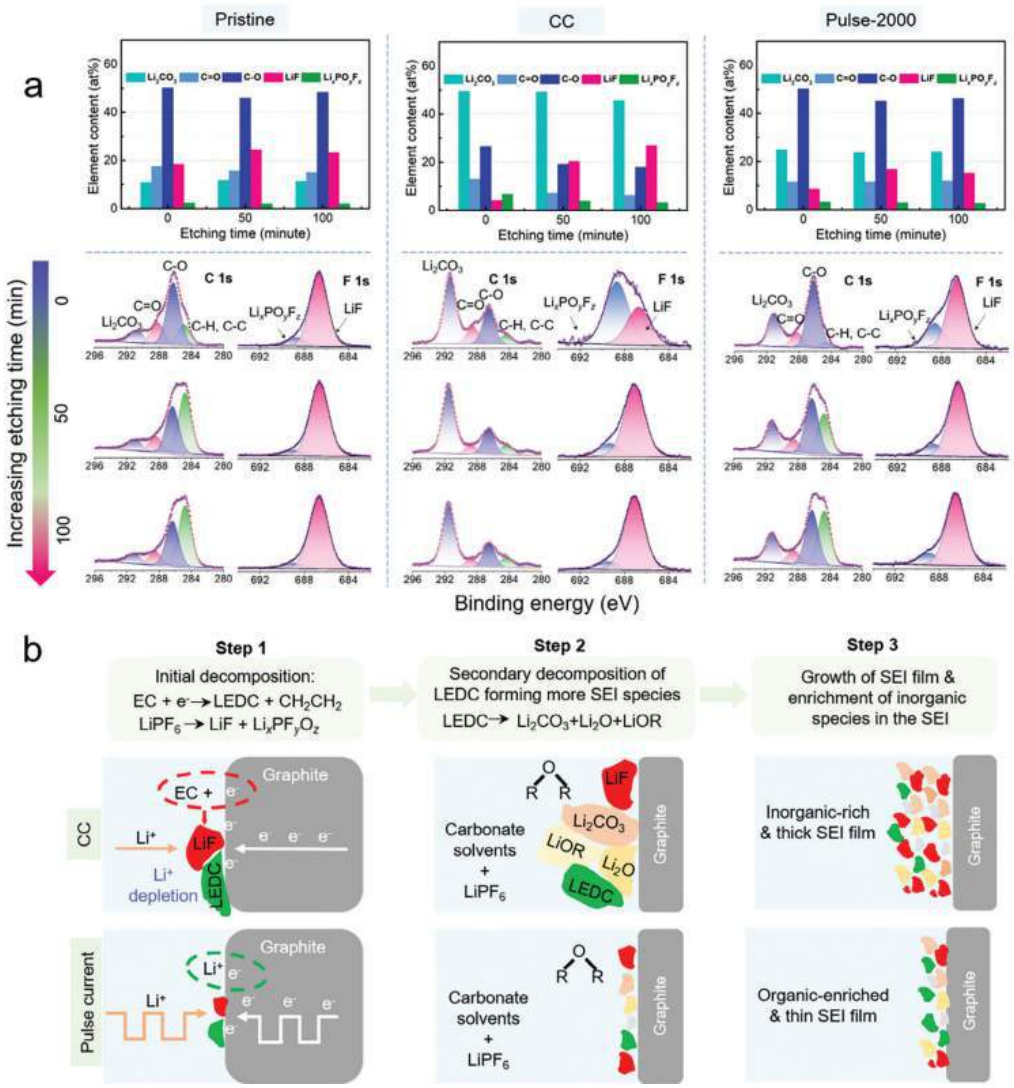


Figure 15: Graphite-electrolyte XPS interface analysis. a) XPS of graphite electrode interfaces: chemical compositions of the SEI of fresh, CC-aged, and Pulse-2000-aged graphite electrodes, with detailed C 1s, F 1s, O 1s, and Li 1s XPS spectra at different etch durations. b) Mechanism of SEI film growth at the graphite interface under constant current and pulse current charging, based on the SEM, EDX, and XPS results above (Guo et al., 2024).

Conclusions

This review highlights advances in understanding, predicting, and reducing lithium-ion battery degradation. Battery science is moving beyond simple capacity metrics toward an integrated approach that combines machine learning, electrochemical diagnostics, materials characterization, and field data. The studies reviewed show that degradation arises from a variety of interacting mechanisms, including loss of lithium inventory, loss of active material, interfacial film growth, electrode damage, and impedance increase. Understanding these interactions is key to improving battery life and reliability.

Early-cycle data alone has been shown to predict long-term battery health for a particular material system and cell design using ML. Voltage-curve features from the first few cycles reveal degradation pathways long before measurable capacity fade occurs. These ML techniques enable earlier and more accurate lifetime prediction. For manufacturers, they offer faster validation, improved yield, and earlier detection of underperforming cells.

This review also challenges the assumption that constant-current laboratory tests represent battery aging in real-world operation. Contrary to assumptions, recent studies show that realistic dynamic discharge profiles can significantly extend cycle life. ML analyses further identify low-frequency current content, peak current, and time spent at high state of charge as major degradation drivers. These results suggest that future qualification standards should incorporate realistic duty cycles.

The studies use mechanistic analysis from post mortems to help explain these observations. Battery aging was shown to be impacted by SEI and CEI growth, lithium inventory loss, electrode cracking, active-material isolation, and rising charge-transfer resistance. Pulse-current charging has emerged as a promising mitigation strategy to aging. It reduces concentration gradients, suppresses side reactions, and helps preserve electrode integrity. Experimental results with Pulse charging show substantially longer cycle life than conventional charging methods.

Together, these findings point to a shift in battery engineering. The goal is no longer just to measure degradation after it occurs in the field. Battery-management systems are evolving into active lifetime-management platforms that could adjust operating conditions based on predicted degradation risk. Although these studies focus on lithium-ion batteries, the same framework can be used for silicon-rich, sodium-ion, lithium-metal, and solid-state technologies. The combination of advanced analytics, materials science, and intelligent battery management strategies provides a pathway toward lowered costs and sustainable electrification.

Acknowledgments

I would like to thank Sanjeev Datta, CTO of Ohmic Labs for his help and guidance with the manuscript.

Works Cited

Guo, J., Jin, S., Sui, X., Huang, X., Xu, Y., Li, Y., Kristensen, P.K., Wang, D., Pedersen, K., Gurevich, L., & Stroe, D.-I. (2023). Unravelling and quantifying the aging processes of commercial $Li(Ni_{0.5}Co_{0.2}Mn_{0.3})O_2$/graphite lithium-ion batteries under constant current cycling. J. Mater. Chem. A, 11, 41. https://doi.org/10.1039/D2TA05960F

Severson, K.A., Attia, P.M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M.H., Aykol, M., Herring, P.K., Fraggedakis, D., Bazant, M.Z., Harris, S.J., Chueh, W.C., & Braatz, R.D. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nature Energy, 4, 540. https://doi.org/10.1038/s41560-019-0356-8

Geslin, A., Xu, L., Ganapathi, D., Moy, K., Chueh, W.C., & Onori, S. (2025). Dynamic cycling enhances battery lifetime. Nature Energy, 10, 172. https://doi.org/10.1038/s41560-024-01675-8

Guo, J., Xu, Y., Exner, M., Huang, X., Li, Y., Liu, Y., Wang, H., Kowal, J., Zhang, Q., Kristensen, P.K., Wang, D., Pedersen, K., Gurevich, L., Stroe, D.-I., & Adelhelm, P. (2024). Unravelling the mechanism of pulse current charging for enhancing the stability of commercial $LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$/graphite lithium-ion batteries. Advanced Energy Materials, 14(22), 2400190. https://doi.org/10.1002/aenm.202400190

Dubarry, M., Truchot, C., & Liaw, B.Y. (2012). Synthesize battery degradation modes via a diagnostic and prognostic model. J. Power Sources, 219, 204. https://doi.org/10.1016/j.jpowsour.2012.07.016

Gent, W.E., Busse, G.M., & House, K.Z. (2022). The predicted persistence of cobalt in lithium-ion batteries. Nat. Energy, 7, 1132. https://doi.org/10.1038/s41560-022-01129-z

Kirkaldy, N., Samieian, M.A., Offer, G.J., Marinescu, M., & Patel, Y. (2022). Lithium-ion battery degradation: Measuring rapid loss of active silicon in silicon-graphite composite electrodes. ACS Appl. Energy Mater., 5, 13367. https://doi.org/10.1021/acsaem.2c02047

Han, X., Lu, L., Zheng, Y., Feng, X., Li, Z., Li, J., & Ouyang, M. (2019). A review on the key issues of the lithium ion battery degradation among the whole life cycle. eTransportation, 1, 100005. https://doi.org/10.1016/j.etran.2019.100005

Zhu, J., Dewi Darma, M.S., Knapp, M., Sørensen, D.R., Heere, M., Fang, Q., Wang, Z., Dai, H., Mereacre, L., Senyshyn, A., Wei, X., & Ehrenberg, H. (2020). Investigation of lithium-ion battery degradation mechanisms by combining differential voltage analysis and alternating current impedance. J. Power Sources, 448, 28. https://doi.org/10.1016/j.jpowsour.2019.227575

Wen, Y., Fan, J., Shi, C., Dai, P., Hong, Y., Wang, R., Wu, L., Zhou, Z., Li, J., Huang, L., & Sun, S.G. (2019). Probing into the working mechanism of Mg versus Co in enhancing the electrochemical performance of P2-type layered composite for sodium-ion batteries. Nano Energy, 60, 162. https://doi.org/10.1016/j.nanoen.2019.02.074

Huang, X., Liu, W., Acharya, A.B., Meng, J., Teodorescu, R., & Stroe, D.I. (2021). Effect of pulsed current on charging performance of lithium-ion batteries. IEEE Trans. Ind. Electron., 69, 10144. https://doi.org/10.1109/TIE.2021.3121726

Guo, J., & Li, W. (2022). Synthesis of single-crystal $LiNi_{0.7}Co_{0.15}Mn_{0.15}O_2$ materials for Li-ion batteries by a sol-gel method. ACS Appl. Energy Mater., 5, 397. https://doi.org/10.1021/acsaem.1c02939

Li, W.H., Ning, Q.L., Xi, X.T., Hou, B.H., Guo, J.Z., Yang, Y., Chen, B., & Wu, X.L. (2018). Highly improved cycling stability of anion de-/intercalation in the graphite cathode for dual-ion batteries. Adv. Mater., 31, 1804766. https://doi.org/10.1002/adma.201804766

Huang, X., Jinhao, M., Liu, W., Ru, F., Duan, C., & Xu, X. (2023). Lithium-ion battery lifetime extension with positive pulsed current charging. IEEE Transactions on Industrial Electronics, 71(1), 484. https://doi.org/10.1109/TIE.2023.3250850

Abdel Monem, M., Trad, K., Omar, N., Hegazy, O., Mantels, B., Mulder, G., Van den Bossche, P., & Van Mierlo, J. (2015). Lithium-ion batteries: Evaluation study of different charging methodologies based on aging process. Appl. Energy, 152, 143. http://dx.doi.org/10.1016/j.apenergy.2015.02.064

Kannan, D.R., & Weatherspoon, M.H. (2021). The effect of pulse charging on commercial lithium cobalt oxide (LCO) battery characteristics. Int. J. Electrochem. Sci., 16, 210453. https://doi.org/10.20964/2021.04.30

Li, S., Wu, Q., Zhang, D., Liu, Z., He, Y., Wang, Z.L., & Sun, C. (2019). Effects of pulse charging on the performances of lithium-ion batteries. Nano Energy, 56, 555. https://doi.org/10.1016/j.nanoen.2018.11.070

Qian, G., Li, Z., Wang, Y., Xie, X., He, Y., Li, J., Zhu, Y., Xie, S., Cheng, Z., Che, H., Shen, Y., Chen, L., Huang, X., Pianetta, P., Ma, Z.F., Liu, Y., & Li, L. (2022). Value-creating upcycling of retired electric vehicle battery cathodes. Cell Reports Phys. Sci., 3, 100741. https://doi.org/10.1016/j.xcrp.2022.100741

Dubarry, M., Svoboda, V., Hwu, R., & Liaw, B.Y. (2006). Incremental capacity analysis and close-to-equilibrium OCV measurements to quantify capacity fade in commercial rechargeable lithium batteries. Electrochem. Solid-State Lett., 9, A454. https://doi.org/10.1149/1.2221767

Guo, J., Li, Y., Meng, J., Pedersen, K., Gurevich, L., & Stroe, D.I. (2022). Understanding the mechanism of capacity increase during early cycling of commercial NMC/graphite lithium-ion batteries. J. Energy Chem., 74, 34. https://doi.org/10.1016/j.jechem.2022.07.005

Iurilli, P., Brivio, C., & Wood, V. (2021). On the use of electrochemical impedance spectroscopy to characterize and model the aging phenomena of lithium-ion batteries: A critical review. J. Power Sources, 505, 229860. https://doi.org/10.1016/j.jpowsour.2021.229860

Asenbauer, J., Eisenmann, T., Kuenzel, M., Kazzazi, A., Chen, Z., & Bresser, D. (2020). The success story of graphite as a lithium-ion anode material—fundamentals, remaining challenges, and recent developments including silicon (oxide) composites. Sustainable Energy Fuels, 4, 5387. https://doi.org/10.1039/D0SE00175A

Ates, M.N., Jia, Q., Shah, A., Busnaina, A., Mukerjee, S., & Abraham, K.M. (2014). Mitigation of layered to spinel conversion of a Li-rich layered metal oxide cathode material for Li-ion batteries. J. Electrochem. Soc., 161, A290. https://doi.org/10.1149/2.040403jes

Yu, J., Wang, J., Long, X., Chen, L., Cao, Q., Wang, J., Qiu, C., Lim, J., & Yang, S. (2021). Formation of FeOOH nanosheets induces substitutional doping of CeO_{2-x} with high-valence Ni for efficient water oxidation. Adv. Energy Mater., 11, 2002731. https://doi.org/10.1002/aenm.202002731

Ruess, R., Ulherr, M.A., Trevisanello, E., Schröder, S., Henss, A., & Janek, J. (2022). Transition metal oxides and Li_2CO_3 as precursors for the synthesis of Ni-rich single-crystalline NCM for sustainable lithium-ion battery production. J. Electrochem. Soc., 169, 070531. https://doi.org/10.1149/1945-7111/ac8242

Lee, W., Lee, S., Lee, E., Choi, M., Thangavel, R., Lee, Y., & Yoon, W.S. (2022). Destabilization of the surface structure of Ni-rich layered materials by water-washing process. Energy Storage Mater., 44, 441. https://doi.org/10.1016/j.ensm.2021.11.006

Heiskanen, S.K., Kim, J., & Lucht, B.L. (2019). Generation and evolution of the solid electrolyte interphase of lithium-ion batteries. Joule, 3(10), 2322. https://doi.org/10.1016/j.joule.2019.08.018

Cicvaric, K., Pohlmann, S., Zhang, B., Rahmanian, F., Merker, L., Gaberscek, M., & Stein, H.S. (2024). Fast formation of anode-free Li-metal batteries by pulsed current. Phys. Chem. Chem. Phys., 26, 14713. https://doi.org/10.1039/D4CP00775A

Guo, J., & Olsen, M. (2026). Pulsed current charging for sodium-ion batteries. Aarhus University. https://pure.au.dk/portal/en/projects/pulsed-current-charging-for-sodium-ion-batteries/

Modeling Exploratory GenAI Exposure Across the Animation Production Pipeline: A Scenario-Based Analysis By Santosh Dasari

Abstract

In the last few years, creative industries as a whole have been affected by the arrival of Artificial Intelligence across all facets of production. This has raised eyebrows on the role of AI in creative production. Existing reports describe the overall disruption within creative industries but only provide a limited analysis of differences across each stage in production, as well as giving a definitive picture to up and coming artists of what they can expect, specifically based on their observations, this is exceedingly important in volatile careers like animation (CVL Economics). This study presents a model founded on published disruption estimations and evidence informed vulnerability weights. Based on industry survey data and task by task analysis, Industry professionals perceived a near equal effect in all skills involved in either stage of production with a slight emphasis in production followed by post-production. This is largely due to the swifter presumption of repetitive tasks that dominate in either sector. This data was then used to construct a visual model using statistical methods like OLS (Ordinary Least Squares), and vulnerability weighting, and simplified sequential modeling to construct a visual model representing preexisting data. It overall showed a very steep increase of involvement speed and disruption considering the path is undisturbed. All of this, however, rides on a caveat of being specific and scenario based. As well as strictly not for forecasting, but rather representing what's currently believed.

Introduction

In January of 2021, OpenAI released DALL-E, a software designed to generate images based on a simple prompt. This technology showed the increasing ability of GenAI to do creative tasks originally associated with humans. What the world wasn't prepared for was the subsequent advancements of this type of technology would bring to all facets of the world in the coming years. Of all the industries affected, creative industries took a considerably large hit. Everything from music, gaming, and film was affected, among these as the eye of the storm was the animation industry. The process was detailed by Shannon Riffe in their article for CMU: "How AI Is Changing Animation Jobs, Workflows, and Creative Control." This is due to how animation productions require a large cost in order to produce, often requiring large teams of animators and other artisans across budgets exceeding one hundred million dollars (Riffe). This produces the content over the course of multiple years. Meaning that a tool that can accelerate that process indefinitely such as artificial intelligence, as demonstrated by emerging text-to-video systems like Sora AI, will most definitely be used in order to expedite the overall process. This process of AI integration isn't restricted to a single task or production stage but is instead throughout the workflow as it is capable of managing tasks from visual, auditory to text based dimensions (CVL Economics). Most Importantly however, due to this ability of AI, the security

of occupations in creative industries, (especially animation) are vulnerable to this new way to expedite the process.

Currently there has been a modest effort to understand the effect of AI in the animation industry. For example, in 2024, The Animation Guild, a union based on protecting and campaigning for the rights of animators for a healthy livelihood commissioned a report, which was titled: *Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs*, from CVL Economics, an economic consulting firm founded in Los Angeles that specialized in modeling research to understand arts/culture and labor markets, in order to assess down to the numbers the scale of AI involvement in creative industries. This revealed a plethora of notable findings, a quality few stating that as of 2023, 68.7% of the industry have adopted GenAI technology to help manage tasks within animation production (CVL Economics). As the metric was from three years prior, seeing prolonged advances in AI ability, adoption levels have morphed since the publication of the report, especially from the advancement of AI technology since then (Eloundou et al. 2), that number has likely gone up. Apart from such, it was also noted that the most affected careers include, 3D modelers, Sound Editors and Concept Artists (CVL Economics). As well as how allocation of the usage of GenAI across production follows, 80% in post-production, 71% in production and 59% in pre-production (CVL Economics). That last relationship will be very important to the final model. Lastly, according to the World Economic Forum, 86% of employers expect AI and information-processing tech to transform their businesses by 2030 (World Economic Forum 7). The report also goes on to estimate that 39% of workers core skills will change by 2030, underscoring the need for constant workforce adaptation (World Economic Forum 33). In essence, the above information from all sources indicated that Artificial intelligence has shaken the industry severely in years past, is currently now and will continue to do so with greater intensity in the years coming forward.

Furthermore, although animation has had its own fair-share of research done within the past few years, it's important to note that there is still a lot left to be desired. For example, relatively few examine how disruption can differ across the major stages of production within the animation pipeline. Along with the fact that animation is oftentimes grouped with the larger entertainment industry, it is difficult to find data that specifically aid up and coming animators in their hunt for a career. This is made more difficult with a lack of transparent studio models that can accurately reflect the current trends and their possible future, which could potentially guide newly graduated animation students into selecting and curating their career path within the industry.

To address this blatant gap within industry resources, this study developed a scenario-based model that estimated how Artificial Intelligence exposure can vary across the different stages of the production pipeline. This was done through the usage of published industry reports, transparent assumptions and pre-existing research. The model in question doesn't intend to predict the future; it merely aims to provide an evidence-informed scenario that it intended to encourage discussion and awareness about the effect of AI on animation workflows as well as how it can be expected to accelerate within future years. All the while presenting the

possible skills that may become more and more valuable for the people looking to enter the animated workforce.

In remainder, this paper will serve to review relevant literature on artificial intelligence, animated workflows, and labor-exposure based on tasks constructing the industry. Next, will be discussion of the study's methodology, followed by a presentation of the results of the scenario model, describing its limitations as well as its implications, and identifying opportunities for future study. To do this, the most important thing to do is to formulate research questions that will inform the results and methodology of the final study. Research Question 1(RQ1), which stages of the animation pipeline appear most susceptible to AI-driven task automation based on the evidence currently in hand? Research Question 2(RQ2), how will a scenario-based model represent AI exposure across pre-production, production, and post-production across a proposed timeline? And lastly, Research Question 3(RQ3), what implications will these findings have for future animation professionals? These questions will guide the presented research.

Literature Review

Artificial Intelligence and Labor

Research regarding GenAI in recent times has been understood to be affecting industries and individual work through the automation of tasks that then disrupt an occupation rather than replacing it entirely. This is supported by Eloundou et al. in "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", which describes how occupational exposure is calculated through the combination of exposure scores of the tasks that define each occupation. This supports the claim due to how it informs that occupational exposure could be appraised through tasks making up each occupation, often leading to affect and therefore disruption, based on exposure ratings on a task by task basis. Therefore, it implies that tasks are the root of disruption, so in the study of such occurrences they should be the focus. Riding off of this, researchers reject the assumption of complete occupational automation and instead shift their focus on evaluating which specific activities AI will perform and how those capabilities influence workflow productivity. This is supported by "Experimental Evidence on the Productivity Effects of Generative AI", in this paper it is found that participants completed writing tasks more quickly and with improved average quality when GenAI assisted them in doing so (Noy and Zhang). This supports how researchers focus on the evaluation of specific activities rather than assuming workplace elimination due to how it shows AI only expedited the process when used in assistance on specific tasks, rather than describing it as deleting the need for human involvement itself.

AI in Creative Industries

In discussion of expedited processes, creative industries represent one of the earliest sectors to show large discussion about AI integration. Increasingly over the times, artificial intelligence is being weaved into ideation, asset generation and support for production, whilst its

degree of adoption varies across disciplines (CVL Economics). The appearance of AI in these industries has long been a spectre to workers, as seen in The Animation Guild (TAG)'s rally held in August 10th of 2024, where 2,500 animators rallied together in Burbank, CA to protest AI job displacement and unpaid job creep ("Thousands Show Support"). The sheer size and notable figures involved (i.e. Rebecca Sugar, James Baxter, Genndy Tartakovsky) displays how imperative the issue had become amongst industry professionals. Specifically for animation, this is further proven by CVL Economics' report, denoting a projected 21.4% of industry jobs disrupted by 2026 (CVL Economics). Roughly one-fifth of industry employment was predicted to experience integration in accordance with the report's estimate, a number that later analysis will prove to only increase, by a steep margin. How this disruption is distributed, however is even mor important to take into account

CVL Economic's report emphasizes the disruption's varied presence across production stages. This report indicates, 80% disruption for post-production, 71% disruption for production and 59% disruption for preproduction (CVL Economics). This statistic, formed based on CVL Economics' survey of 300 industry professionals, displays crucial information for young animators heading into the industry. It shows how occupations within post-production are almost all affected, followed closely by the other two stages, roughly only 10 percent apart. This statistic gives effective clues to answer the second research question, (RQ2): How will a scenario-based model represent AI exposure across pre-production, production, and post-production across a proposed timeline? In short, a heavy majority across all sectors, yet skewing towards the end of the pipeline. This information will be useful in mapping vulnerability weights to construct a visual representation of this report later on (the aforementioned model).

Animation Pipeline

The raw data for this is surely revealing, but in order to thoroughly understand this, however, one must come to know how the animation pipeline itself works. In *The Animation Field Guide*, Epic Games describes the evolution of workflows in animation production through three main pipelines: a) Traditional 2D/cel, b) Traditional 3D, and c) real-time engine-based pipelines (12-14). A) follows a strict linear sequence, this pipeline was the foundation for all animated productions dating back to the early 20th century. Starting from a script, then building off to storyboards, to hand drawn frames, refined through ink and paint, photography and finally assembly, where early design decisions remain fixed to avoid costly re-draws (12). It is important to note the storyboarders themselves use working character/prop/environment designs, made by the designers and explored by the concept artists whilst they are finalized along the way throughout production. This process has since evolved into the digital age where hand drawn frames are imprinted into a tablet instead, taking the storyboard-churned layout designs into the animation process itself. From there the same fundamental animation principles stay in touch, starting with the main storytelling of key-frames, defining the precipice of the movement with extremes, and then filling the gaps through the usage of breakdowns/inbetweens. This is then revised and given to a colorist. Modern assembly, at least before the onset of artificial

intelligence, includes compositing all of the clips together, piercing together all of the separate animations in one scene, along with the background, score, and lighting/editing effects to bring the entire clip to life. B) Epic Games notes the 3D process to function more similarly to a linear siloed structure, where assets pass through pre-vis, layout, rigging and keyframe animation, lighting and then final assembly. A major bottleneck described in this process however is the lengthy times needed for offline rendering and compositing making changes downstream excessively time-consuming (13). C) the final pipeline depicted is based on real-time engines, where the aforementioned bottleneck is replaced with a parallel, non-linear setting that is charged by a game engine, such as unreal engine. By using real-time rendering, departments like animation, lighting and layout work simultaneously with instant visual feedback. This allows creators to work continuously, test dynamic camera angles and finalize pixels in realtime insteading waiting for the render farm(15).

With the pipeline established, pre-existing research on exposure to AI can be connected to specific production flows. The research has explained clearly that Artificial Intelligence's effect stems from how it can automate tasks rather than it completely getting rid of occupations in general. It's also been discussed how its effect is distributed across the production pipeline. However, few studies have provided a visual model to describe this latter finding. Sure, it was seen in the CVL Economics report that AI disruption is weighted towards the latter parts of production (post-production>production>pre-production), however, what isn't available is a visual model representing the rate at which this is happening as presented by reports such as the *Future Unscripted b*y CVL Economics. In doing so, all three of the guiding research questions can be answered. That is, with a prior explanation of the methods employed in order to create such a model.

Methodology

Study Design

It should be very clear what this model is in order to avoid confusion when making claims based off of it. This study created a scenario-based quantitative analysis that visualizes preexisting labor data, to estimate how AI exposure may vary across stages of animation production. What this study does NOT do, is predict future employment outcomes. This is due to the fact that it is not crafted to establish statistical significance as the model is based on scenario-assumptions rather than longitudinal studies. Despite that fact, it can still prove useful for up and coming animators, as it effectively illustrates a temporal evidence-informed scenario/trend formed from current data that shows the acceleration of AI disruption across the production pipeline. The visual representation itself is consolidated within a simple regression model. The dependent variable was the production stage's allocated disruption percentage, and the independent variable was a time value, as the numbers went up in the x-axis, more time was projected to pass by. The AI disruption/exposure rate was handled cumulatively, meaning that once the vulnerability rates were assigned, and each stage of production had its own dedicated

fraction of the overall 21.4% predicted exposure, the values become added up at each next data point/production stage at its own dedicated time value until the total 21.4% is reached. This creates a smooth linear regression, whose slope shows the acceleration of AI adoption conceived from CVL Economic's report.

Data Sources

This study used a combination of peer-reviewed research articles, industry reports and workforce analyses made publicly available to create a scenario based model of potential AI exposure across the animation production pipeline. As no single source provides a stage-specific acceleration model of AI disruption within animation, multiple sources were synthesized to establish a transparent framework informed by evidence. The primary source used in this study is *Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs*. This provided survey data and assessments of the adoption of GenAI across animation/entertainment production tasks. This report was used to ID AI usages within each production stage, and to inform the relative exposure assumptions assigned to each stage.

In order to establish the foundation for an analysis based task-for-task, this study draws upon Eloundou et al.'s "GPTs are GPTs: An Early look at the Labor Market Potential of Large Models". Instead of evaluating occupations as affected uniformly, this study measures disruption caused by artificial intelligence at the task level, this framework then supports the choice to analyze animation as a collection of production stages constructed from different task groups.

Extra context regarding workforce transformation and evolving skill needs was attained from the World Economic Forum's *Future of Jobs Report 2025*. This report was used to provide more general evidence on AI adoption, anticipated workforce changes in skill demands, and expectations for employers surrounding emerging tech.

Economic and technological context was further supported with McKinsey's *The Economic Potential of Generative AI: The Next Productivity Frontier*. The report provides evidence in regards to the productivity of GenAI and the potential for it to augment work across industries, and support the rationale for examining the influence of AI on creative workflow.

Moving on, Noy and Zhang's experimental study on the productivity effects of GenAI technology was implemented to show how artificial intelligence can improve the effectiveness and efficiency of tasks. This study provides a direct complement to the task-based framework by showing trackable gains in performance, when GenAI is applied to knowledge-intensive activities.

In essence, all of this material provides complementary evidence that supports the construction of a scenario-based transparent model. Each of the sources address a different aspect of AI adoption, the combined use enables the study to evaluate potential differences in the exposure of Artificial Intelligence across the pipeline, in signature task-for-task basis, whilst acknowledging the mental haze that accompanies the forecasting of emerging tech.

Evidence Backed Assumptions

In lieu of the lack of substantial data points to conceive from a personal survey, due to how hard it is to obtain a large representative survey sample in the given timeframe, this study uses evidence-backed assumptions from existing reports from industry. First of all, the overall disruption estimate was assumed to be 21.4%, this figure was previously mentioned to have been a part of the CVL Economics report. Now the assumption part of this comes from the fact that this study revolves around the fact that the previous figure was for the estimated disruption across film, television AND animation. This assumption, or rather, simplification, is acceptable as all of the AI-exposed tasks whose disruption was computed to calculate the overall figure are found in the animation pipeline as well as for TV and film. And since the CVL Economics report group animation occupations in this category in the first place that implies that using the overall estimate and modeling a regression off of such will remain helpful to animators, the only extraneous variable being that it offer help to other careers as well, which in this context can't be bad thing.

Next up, the second assumption taken was how the timeline represents one plausible scenario. What this means is, that based on the current patterns observed within the CVL Economics report, the model displays an extended visualization of the projection echoed throughout the writing. And since there is a limited amount of variables on the regression, it only accounts for a linear visualization, not taking into account any extraneous/hindering effects that may take place as the AI age moves on, that may very well be set off course. However as of now, using this one specific scenario will prove efficient as per the reliability of an official economic firm's projection.

Lastly in the same sense as the last assumption, the model's x values of time progression, aren't uniformly established units of time such as a year, or decade or what not. They are arbitrary values to represent AI adoption from quickest to slowest. Keeping all of the above mentioned in mind it is now time to show how the model itself was conceived.

Model Construction

What Is Sequential Modeling

The design needed to visualize the acceleration of AI disruption of the future was born from the theory of sequential modeling. The Moveworks AI Terms Glossary defines sequential modeling as a process that uses chronologically ordered data to underscore patterns and identify patterns over the times (Moveworks). Though this study didn't use a trained machine learning algorithm, it applied the same fundamental concept. Through combining a chronological projection (21.4% industry job disruption by 2026) in union with task-originated vulnerability rates, the dataset was constructed sequentially. This then let the regression model build upon evidence-backed assumptions to then visualize the acceleration of disruption within the coming years.

Sector Sorting

In order to craft the points of data required to conceive the regression model, as corroborated by Eloundou et al. in *GPTs are GPTs*, the disruption based on task level had to be analyzed and ordered to the organization design of the dependent variable (production stage). Luckily, CVL Economics' report included this exact metric. And due to how it was this task based data that informed the 21.4% metric in the CVL Economics report, this means how the tasks are split into either of the three production stages will inform the rate this 21.4% prediction will accumulate (ex. decreased over time, increased over time or constant) based on which production stage is proven to be most to least susceptible. The tasks that needed to be organized also came with their own unique disruption rate which was perfect to use in order to create the order in which production stage was more susceptible than the other (by seeing which stage of tasks had the higher average GenAI impact). Below is a table on the tasks that were organized in question:

Task	**% of respondents reported Impact/Disruption**	**Sample Jobs**
3D Modeling	44%	3D Modeler, CG modeler, Design Engineer, Product Design Manager, Video Designer, Motion graphic artist
Character and Environmental Design	39%	Illustrator, Concept Artist, Environmental Artist, Character Artist, Cartoonist
Voice Generation and Cloning	37%	Sound Designer, Sound Editor, Mix Engineer, Music Editor
Compositing	37%	Compositor, Nuke Compositor, Motion Designer, and FX Technical DIrector
Sound Design	34%	Sound Editor, Sound Designer, Re-recording Mixer
Tools Programming	29%	Digital Interface Designer, Broadcast Technician, Software Engineer, Technical Project Manager, Technical Artist
Script Writing	29%	Script writer, Associate Producer, Production Assistant
Animation and Rigging	27%	Special Effects Artist, Animator, Graphic Designer, Technical Animator, Rigging Manager, Entertainment Technician
Concept Art/ Visual Development	26%	Storyboard Artist, Concept Artist, Creative Director, Graphic Designer
Light/Texture Generation	25%	Texture Artist, Look/Surface/Materials Artist, Background Painter, Environment Artist, Modeler, Lighting Technician

Table 1: GenAI Impact on FIlm, Television, and Animation Industry Tasks by CVL Economics

These tasks were then organized into preproduction, production and postproduction, using the previously mentioned information on the animation pipeline.

Pre-production	Production	Post-production
Script Writing	3D Modeling	Light and Texture Generation
Concept Art/Visual Development	Animation/Rigging	Compositing
Character Design		Voice Generation & Cloning
		Sound Design

Table 2: Occupations organized into separate production stage

Vulnerability Mapping

Once these were sorted, the averages of the task disruption per group were performed. This resulted in averages of 31.3% for pre-production, 33.3% for production, and 33.25% for post-production. Similar to the reports of GenAI use in the industry, production and post-production were the highest, earlier at the task by task basis, production was slightly higher. The close similarity between production and post-production reflects the limited difference in task-level exposure reported across stages rather than a large vulnerability separation. The reason why this was used instead of the broad metric of production stage by stage, which states post-production as highest (all separated by larger margins). This is due to the aforementioned article by Eloundou et al. which noted the importance of assessing AI disruption in terms of task, therefore when modeling its potential acceleration, it's important to keep that in mind. From these average percentages corresponding vulnerability weights were added to display the weighting of each, the key was to make them barely different yet enough to be noticeable to procure a model, as that is what the data communicated. Vulnerability weights are decimal values that when added in total have to equal one, and when multiplied by the overall percentage(in this case 21.4%), it gives the percentage share of that corresponding sector. The vulnerability rates for each stage of production(the sectors used) were 0.29 for pre-production, 0.36 for production and 0.35 for post-production.

Production Stage	Sector's Percentage share
Pre-Production	6.206%
Production	7.704%
Post-Production	7.49%

Table 3: Each production stage's disruption share

Above are the percentage share each sector will now carry in terms of the final regression. Next using the aforementioned process described in the section on study design. In recap, the time values of 1-3 (quickest to slowest) adoption of AI disruption will be assigned to

each production stage in order of how quickly either has proven through the task analysis to be disrupted. The y-values will be the cumulative total stage by stage of each production riding until that 2026 projection deadline made by CVL Economics.

X (Time)	Y (% of industry disruption completed)
1	7.704 (Production)
2	15.194 (Production + Post-Production)
3	21.4 (Production + Post-Production + Pre-Production)

Table 4: Final data points for regression

Using Ordinary Least Squares (OLS)

The consequential slope and graph made by using software such as Desmos conceive the model the paper has been building up to. However, one last thing before the results are revealed, the slope for this, as it is trying to find the line of best fit, is designed through the Ordinary Least Squares AKA OLS. This process functions by calculating the slope and y-intercept by minimizing the sum of the squared vertical distances(residuals) in between the data points and the line ("Nonlinear Regressions"). The below formula brings this about, and it is what the desmos application refers to when constructing the line of best based on imputed points.

First the variables should all be named to understand how the calculation in the formula(s) works:

Variable	Definition
m	slope
b	y-intercept
Σ	Summation operator (Capital Sigma)
X_i, Y_i	Individual pair of x and y values from dataset
$\overline{X}$	Mean of all x-values in data set
$\overline{Y}$	Mean of all y-values in data set

Table 5: OLS Equation variables

Ordinary Least Squares (OLS) Formula(s)

$$Slope\ (m) = \frac{\Sigma(Xi - \bar{X})(Yi - \bar{Y})}{\Sigma(Xi - \bar{X})^2}$$

Y-intercept (b) = $\bar{Y}$ - $m\bar{x}$

Once calculated, this results in a slope(*m*) of 6.848 and a y-intercept(*b*) of 1.07. The results section will display the final regression, the special Desmos terminology and what this means overall.

Results

Desmos Variable	Definition
R^2	Coefficient of Determination, measures how well the regression line fits the data.
r	Pearson Correlation Coefficient, measures the strength and aim of the linear relationship ("What is RMSE")
X_i,Y_i	Table Headers

Table 6: Desmos variable definitions

With the variables finally defined here is the final model

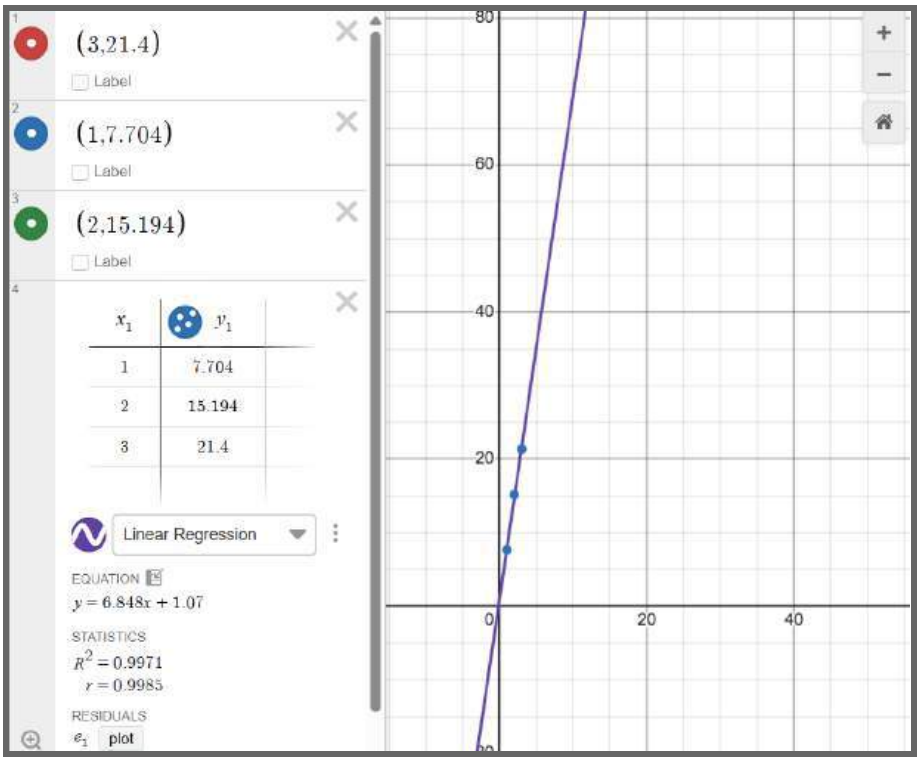


Fig 1: Final Regression of potential future disruption acceleration

Keep in mind, before dissection, though desmos doesn't have axis titles, the X-axis is time, and the Y-axis is the percentage of industry disrupted. Moving on, at face level the values on the left side say a lot, the coefficient of determination is ranged from 0 to 1, therefore a value of 0.9971, or 99.71% indicates an extremely strong fit within the data set. The Pearson

Correlation Coefficient ranges from -1 to +1, therefore a value of 0.9985 shows a strong positive linear relationship. Overall these numbers confirm that the regression closely shows the relationship in the dataset.

When the graph is analyzed, it shows a direct relationship between the variables and a steep linear regression that increases over time. Although as previously mentioned, no specific times can be assumed, what this graph does provide is a visualization of the expert data designed by CVL Economics in their study. It illustrates the trend they were describing, in how it shows a scenario of increasing cumulative exposure under the assumptions made, claiming (in its own designated amount) production, post-production, and pre-production in order by 2026 for that mentioned total of 21.4% and then rapidly accelerating beyond that.

For this paper that is the claim it will rest at, however interestingly, when quadratic regression is used the points line up similarly. This could suggest that at a certain point, the jobs taken will plateau (be completely done), and then AI disruption decreases because there's nothing to disrupt, however the research done, specifically CVL Economics directly denies this. However it is an interesting development, in that it also fits the points nicely.

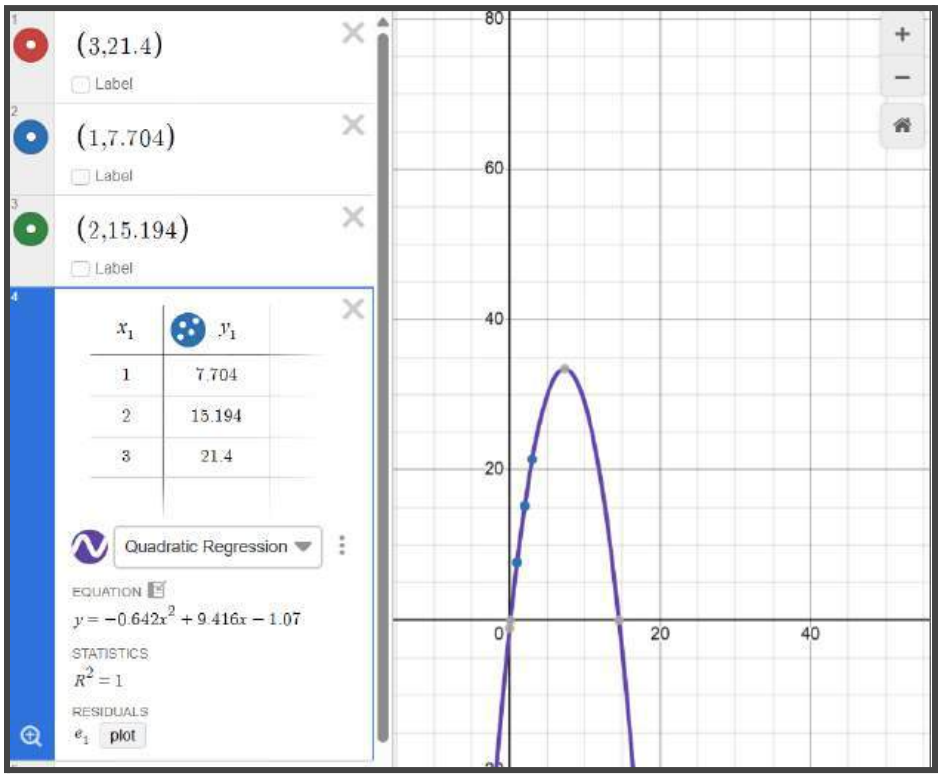


Fig 2: Quadratic regression of previous data points

One last thing should be mentioned before the conclusion however, there is a way this can go wrong. As initially attempted, a non-cumulative approach can be done for the y-values, which leads to regression (using OLS) that looks more like this:

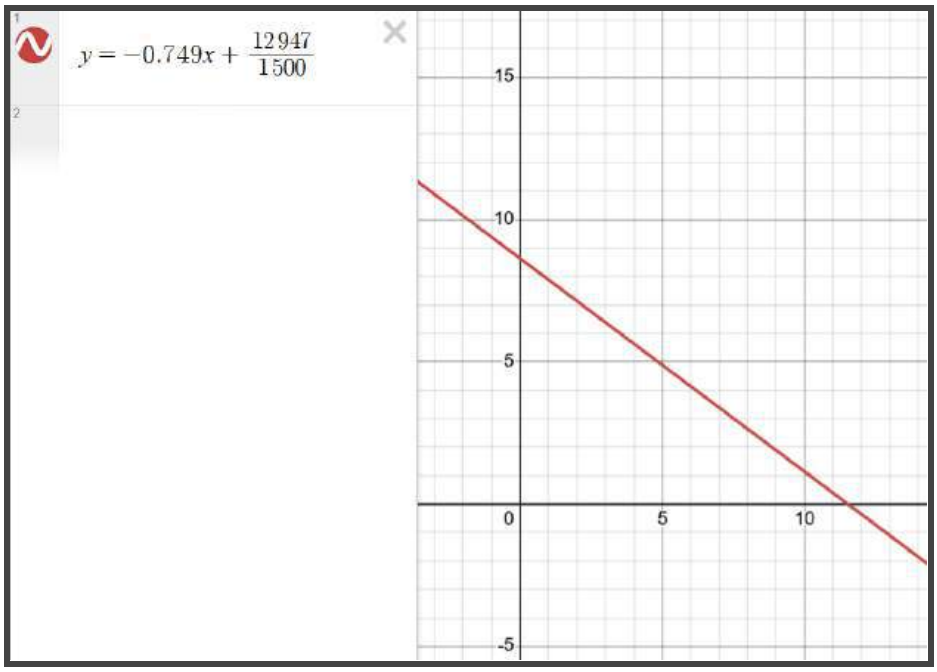


Fig 3: Initial error regression resulting from misinterpretation of data points

It's important to note that this would be incorrect because it disregards and discards the disruption in the stages before and implies AI disruption will decrease, which from the CVL report is factually incorrect. Therefore this should be noted for anyone looking to replicate the study.

Questions Answered

Based on the model and supporting evidence, the research questions can be addressed. In recap the questions were as follows: Research Question 1(RQ1), Which stages of the animation pipeline appear most susceptible to AI-driven task automation based on the evidence currently in hand? Research Question 2(RQ2) How will a scenario-based model represent AI exposure across pre-production, production, and post-production across a proposed timeline? And lastly, Research Question 3(RQ3), What implications will these findings have for future animation professionals? RQ1 is effectively answered as the most susceptible stage of production according to the current evidence is production followed closely by post-production, all three being in close contention; however. RQ2 is answered through how the model crafted represented it as a steeply upward sloping line, showing assumption of quickly adopted technology. RQ3 is answered through this data implies that animators should be aware of increased and more rapid AI involvement in the coming years, prioritizing not only understanding AI in pipeline, but sharpening their pre-production skills, as the model describes such as slowest to be adopted, albeit by a close margin.

This implies that apart from sharpening pre-production skills, individuals looking into entering the industry should pay close attention to their own ability to create complex and novel ideas while relying on the unguided reliance on AI-generated outputs that is using AI to create

art. That sentiment is supported by the World Economic Forum's statement on the priority importance of creative thinking and how AI is one of the fastest growing skill areas (World Economic Forum). The aforementioned unguided process suggestion is backed by Doshi and Hauser's claim on how AI-assisted creations lose diversity over time (Doshi and Hauser). As well as closely studying how AI is evolving beyond the acceleration projected to map the directions to follow.

Conclusion

All told, this study developed a scenario-based transparent model, a scenario describing current trends, to understand how GenAI can affect the various stages of the animation pipeline. By consolidating peer-reviewed research, as well as industry reports, differences in task composition were explored in how that may lead to differing levels of exposure across preproduction, production. And postproduction in terms of AI involvement.

Once again, it cannot be stressed enough, this study is meant to transparently model a single scenario, not be used as a forecasting device. The model is evidence informed to every detail as possibly found in the sources that was applicable, meaning it can be a very helpful resource for up and coming animators to visualize the future based on currently envisioned trends rather than definitely predicting the future.

All the while, it demonstrates the value of task-based understanding for appraising emerging tech in creative industries especially animation. The acceleration model gives a Pearson correlation coefficient and coefficient of determination-verified regression that shows how steep and therefore dramatic AI involvement will continue to be. It incites a prioritization in production for upcoming AI disruption followed by post-production closely, whilst proving that future animators should prioritize pre-production skills whilst knowing since difference is only by a strict and meager margin between production stages, that learning how these AI tools work and developing the ability to make original content and direct AI-workflows well is important, And as artificial intelligence's abilities evolve, research that implements empirical employment data, longitudinal analyses, and industry surveys will be needed to refine the estimates and to understand their future implications more effectively for animation productions.

Works Cited

Animation Magazine. “Thousands Show Support at TAG Rally Against AI and Unfair Labor.” Animation Magazine, 10 Aug. 2024.

CVL Economics. Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs. The Animation Guild, Jan 2024.

Desmos. “Nonlinear Regressions.” Desmos Help Center, https://support.desmos.com/hc/en-us/articles/360042428612-Nonlinear-Regressions. Accessed 31 July 2026.

—. “What is RMSE?” Desmos Help Center, https://help.desmos.com/hc/en-us/articles/202529129-What-is-RMSE. Accessed 31 July 2026

Doshi, Anil R., and Oliver P. Hauser. “Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content.” Science Advances, vol. 10, no.28, 2024, p. Ead5290.

Eloundou, Tyna, et al. “GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.” arXiv, 2023, https://arxiv.org/abs/2303.10130

Epic Games. The Animation Field Guide. Epic Games, 2022.

McKinsey & Company. The Economic Potential of Generative AI: The Next Productivity Frontier. McKinsey Global Institute, June 2023

Moveworks. “Sequential Modeling.” Moveworks, moveworks.com. Accessed 31 July 2026.

Noy, Shakked, and Whitney Zhang. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.” Science, vol.381, no.6654, 2023, pp. 187-192

Riffe, Shannon. “How AI Is Changing Animation Jobs, Workflows, and Creative Control.” Carnegie Mellon News, 8 Apr. 2026, etc.cmu.edu/about/news-events/ai-animation-jobs.

World Economic Forum. Future of Jobs Report 2025. World Economic Forum, 2025.

How Does Changing the Axis Ratio of a Logarithmic Potential System Affect Its Integrability? By Muzhi Fang

Abstract

This paper examines the transition from predominantly regular motion to chaotic orbital behaviour through a two-dimensional logarithmic potential, a model used often in studying galactic mechanics, where the axis ratio, q, controls the shape of the potential. As the axis ratio is moved away from q=1 (i.e. circular potential), the y axis of the potential gets squished, breaking rotational symmetry and conservation of angular momentum. The orbits are restricted to a fixed energy surface within the four-dimensional phase space, which is observed qualitatively using a Poincaré section and quantitatively through the Lyapunov exponent calculated using a renormalization method. At q=1 the system is integrable, shown by both the nested loops of the Poincaré section and the MLE convergence towards 0. As q is lowered, the system becomes unintegrable, as seen in the chaotic regions of the Poincaré section as well as a positive MLE; this trend does not hold, however, as surprisingly the MLE peaks at 1.002 around q=0.6 before the system returns to regular motion near q=0.45 and the MLE drops down to 0 once again.

Introduction

Historically, one of the more well-known systems in physics is the two-body system, involving two bodies, behaving as point masses, that orbit each other, being described by Kepler's laws. When Newton later explained these laws using his laws of motion and law of gravitation, it became one of the first systems to be fully solved using physics. While many had assumed that all systems can be solved with enough mathematics, that is not the case. What makes Kepler's problem solvable is a characteristic called integrability. The Liouville-Arnold Theorem was created much later in 1953 by Joseph Liouville and formally proven in by Vladmir Arnold in his book *Mathematical Methods of Classical Mechanics*; a copy of which was later translated into English in 1989. "Liouville proved that if, in a system with n degrees of freedom (i.e., with a 2n-dimensional phase space), n independent first integrals in involution are known, then the system is integrable by quadratures" (Arnol′d, V. I., et al 272).

To better understand integrability, the origins of first integrals (often called conserved quantities in physics) needed to make a system integrable need to be discussed. According to Noether's Theorem each conservation law can be derived from a respective symmetry within the system (Noether). The two conservation laws used in this paper are the conservation of energy and conservation of angular momentum. The conservation of energy is derived from time translation symmetry. Time translation symmetry applies for a system as long as there is no explicit time dependence. In other words, as long as the system does not get affected by moving through time (ie. conducting the same experiment at another point in time does not affect the outcome). The angular momentum conservation results from rotational symmetry, which applies for systems if rotating the system does not affect the physics.

One of the most important things to consider when exploring integrability is chaos. Chaos

can be seen in many cases when a system is non-integrable, which occurs when there are more degrees of freedom than conserved quantities. A classic sign of chaos is its sensitivity to initial conditions (Lorentz). This is seen when 2 systems with extremely close initial conditions quickly deviate until they soon appear to be very different. Henri Poincaré introduced Poincaré sections in 1890 as a way of studying chaotic systems by creating lower dimensional visualizations of high dimensional phase spaces. Another tool used to study chaos is the Lyapunov exponent first used by Aleksandr Lyapunov in his book, which was only later translated into English in 1992, *General Problem of the Stability of Motion* (Lyapunov, A. M).

The main focus of this paper is on a 2D logarithmic potential system, described in the paper "Simple Tools to Study Global Dynamics in Non-Axisymmetric Galactic Potentials – I" (Cincotta, P. M., and C. Simó). Specifically, the goal is to study the trend of the integrability of the logarithmic potential system as it loses a conserved quantity and gets further perturbed from the perfectly integrable system. This will be done both quantitatively and qualitatively through the Lyapunov exponent and the Poincaré section.

In the methods section, the logarithmic potential system used for research is derived. Additionally in the same section, Poincaré sections and Lyapunov exponents, the two methods of showing the results, are explained. Both the Poincaré sections and Lyapunov exponent graphs for integrable and non-integrable versions of the logarithmic potential system are shown and discussed in the results section.

Methods

The Logarithmic Potential System

The system that this paper explores in detail is a 2D logarithmic potential system. While the regular two-body system consists of two point masses orbiting each other, the logarithmic potential system consists of a test mass moving in a potential that is logarithmic as opposed to the regular 1/r potential used for Keplerian systems. One test mass means that this system has 2 degrees of freedom, x and y, and the conservation of energy and angular momentum to make the most basic case of the logarithmic potential system integrable. In galactic mechanics, logarithmic potential systems are often used to study the motion of stars in galaxies; in the real world the speed of stars near the edge of galaxies stays nearly constant as opposed to dropping off significantly as a 1/r potential. To get this behaviour, we need a potential where the circular speed, v_c, stops depending on r once r gets very large. For a circular potential, we can start with:

$$v(r) = v_c$$

This v_c value needs to stay nearly constant at large values of r. Next, the force from the potential is purely radial in the case that the potential is circular. Moving outward by small value dr changes the potential by:

$$dV = \frac{dV}{dr} dr$$

This equals the work done by the force per unit mass from the potential, $a_r dr$ (force per unit mass times displacement); since mass is just a multiplier everything here is per unit mass. Equating the work and the loss in potential gives:

$$a_r = \frac{-dV}{dr}$$

From kinematics, the radial acceleration is equal to

$$a_r = -\frac{v^2}{r}$$

Equating these and plugging in the v_c for the constant speed near the outside of the potential gives:

$$\frac{v_c^2}{r} = \frac{dV}{dr}$$

Integrating the left side:

$$v_c^2 \int \frac{dr}{r} = v_c^2 \ln r + C$$

We ignore the constant for now since it later disappears regardless when differentiating the potential for the EOMs. The important thing is the rate of change of the potential, not the true value. Rearranging for the potential after integrating both sides gives:

$$V(r) = v_c^2 \ln r$$

With the Pythagorean theorem, we can easily see that:

$$r^2 = y^2 + x^2$$

The potential's shape can be easily changed by adding an axis ratio, q, where q=1 would give a perfectly circular potential while q values not equal to 1 will result in ellipses. The axis ratio can be altered to break the conservation of angular momentum later on to explore the behavior of this system when it is no longer integrable. The rotational symmetry of this system falls apart when the $q < 1$ due to the system not being perfectly circular anymore, meaning that the forces on the test mass from the potential will not be solely radial anymore in most cases.

$$r^2 = x^2 + \frac{y^2}{q^2}$$

Assembling the full potential after using log laws to substitute in the term for r squared gives:

$$V(x,y) = \frac{1}{2}v_c^2 \ln\left(x^2 + \frac{y^2}{q^2} + R_c^2\right)$$

Additionally, a new term, the core radius, needs to be introduced for this system. In a regular two-body system, both the potential and the force blows up as the radius approaches 0. This causes problems in simulations for both the Poincaré sections and the Lyapunov exponent after attempting to create these for the two-body system. The core radius for the logarithmic potential system softens the potential near the center and ensures that the potential stays finite even at the origin.

$$V(0,0) = \frac{1}{2}v_c^2 \ln(R_c^2)$$

The equations of motion can be derived directly from the potential. First a new term is defined, since the inside of the logarithm never needs to split:

$$D(x,y) \equiv x^2 + \frac{y^2}{q^2} + R_c^2$$

We differentiate the potential with respect to x, treating y as a constant. The derivative of ln is 1/D. The derivative of D is:

$$\frac{\partial D}{\partial x} = \frac{\partial}{\partial x}(x^2 + \frac{y^2}{q^2} + R_c^2) = 2x$$

Putting it all together:

$$\frac{\partial V}{\partial x} = \frac{1}{2}v_c^2 \cdot \frac{1}{D} \cdot 2x = \frac{v_c^2 x}{D}$$

Let ϕ be the true potential energy of the system (kg m^2 s^{-2}), and V for our logarithmic potential system is potential per unit mass (m^2 s^{-2}); this is the unit of the v_c^2 on the outside of the natural log, and the units inside the log need to be cancelled by dividing all 3 terms with a constant with the units m^{-2}, since logs only take dimensionless calculations. Adding this constant does not affect the EOMs since constants disappear after differentiating anyways. The true potential energy and our logarithmic potential are related by:

$$\phi = mV$$

Conservative forces from a potential can be written as

$$F = -\nabla\phi$$

Where ∇ is the gradient, pointing towards the steepest increase in potential energy. The negative sign flips this, since the force is pulling towards the point where the potential decreases the fastest. Now using Newton's second law and substituting our potential per unit mass for the true potential gives us:

$$m\vec{a} = -m\nabla V \implies \vec{a} = -\nabla V$$

This equation can be split into the two separate components:

$$a_x = -\frac{\partial V}{\partial x} \qquad a_y = -\frac{\partial V}{\partial y}$$

We find the acceleration along the x direction as:

$$\ddot{x} = -\frac{v_c^2 x}{D}$$

Applying the same steps, but differentiating with respect to y gives:

$$\ddot{y} = -\frac{v_c^2 y}{q^2 D}$$

The integrator used for the simulations, RK5(4) takes first order systems, so we use v_x and v_y to avoid the second order equations above. This gives us the following four equations for the integrator.

$$\dot{x} = v_x, \dot{v_x} = -\frac{v_c^2 x}{D} \qquad \dot{y} = v_y, \dot{v_y} = -\frac{v_c^2 y}{q^2 D}$$

These are the 4 coupled first order ODEs that the integrator will be simulating for both the Poincaré sections and the Lyapunov exponents.

Poincaré Sections

Poincaré sections are the main method used in this paper to observe trends in the system as q changes qualitatively. For systems with n variables, the phase space exists in an n dimensional phase space. Studying the phase space in high dimensions is extremely confusing to be worked with mathematically and unable to be processed by our brain which only goes up to 3D. A Poincaré section solves this issue by creating a 2D representation of this phase space. A 2D surface is chosen in this phase space. Now, instead of following the continuous path of the system through this phase space, we only mark the points where the trajectory crosses this section in a certain direction.

In the case of this paper, the logarithmic potential system has 4 variables (x, y, v_x, v_y), giving it a 4 dimensional phase space, which is unable to be visualised, requiring a Poincaré

section. The process to create a Poincaré sections for this system is:

- Since energy is conserved in the system, the last variable can always be solved when the other 3 are known. Therefore, setting a constant value for energy restricts the system to a 3D surface within the 4D phase space. The total specific energy of the system, kinetic plus potential, which is the Hamiltonian H(x, y, v_x, v_y) = E, is expressed by:

$$E = \frac{1}{2}(v_x^2 + v_y^2) + \frac{1}{2}v_c^2 \ln\left(x^2 + \frac{y^2}{q^2} + R_c^2\right)$$

- The next reduction comes from choosing a Poincaré section by fixing one coordinate; the section used for simulations in this paper is y=0, marking the x and v_x for all crossings where $v_y > 0$.

Poincaré sections make it easy to spot periodic orbits in contrast to chaotic ones. Fixed points, usually at the center of nested loops, are from the trajectory passing through every time it crosses the section representing a closed orbit in physical space, since it reaches the same state every period. Closed loops on the Poincaré section are formed by quasiperiodic orbits that never intersect themselves but staying restricted to an invariant torus, a doughnut-like surface in the phase space that the Poincaré section slices, leaving the loop shape on the Poincaré section. The third shape to appear on the section is scattered dots. These are only seen in chaotic systems, when orbits are destroyed leaving dots scattered around the section instead.

An interesting phenomenon is the closed loops that continue to exist on chaotic Poincaré sections. This phenomenon can be explained by the Kolmogorov–Arnold–Moser (KAM) theorem, which states that for many quasiperiodic orbits, slight perturbations do not completely destroy them, only distorting the shapes of some of the tori (Kolmogorov, 1954). It is only when a significant perturbation is added to the system that these stronger tori break.

Lyapunov Exponents

The Lyapunov exponent, λ, is a great measure of chaos within systems since it provides a quantitative way of detecting chaos instead of purely visually, like the Poincaré sections (Lyapunov, 1992). One of the main concepts for chaos is its sensitivity to initial conditions. Specifically, when a system is chaotic, two initial conditions starting a tiny distance apart drift apart quickly. The Lyapunov exponent measures how fast this happens. We compare pairs of trajectories that start at nearly the same but slightly different positions. Instead of the [x, y, v_x, v_y] of the main trajectory, the second one starts at [x+0.001, y, v_x, v_y] and the path difference is tracked. This initial separation is denoted as δ_0. For chaotic trajectories, the separation grows according to:

$$\delta(t) \approx \delta_0 e^{\lambda t}$$

Where the lyapunov exponent used is the maximal lyapunov exponent (MLE) in the system. A max value is needed because in some systems a lot of orbits are still stable and only

the Lyapunov exponents of some orbit pairs are matching that of a chaotic orbit. Solving this relationship for λ gives:

$$\lambda = \lim_{t \to \text{inf}} \frac{1}{t} \ln\left(\frac{\delta(t)}{\delta_0}\right)$$

Looking at this relationship shows that only if the separation grows exponentially, the sign of a chaotic orbit, will the lyapunov exponent be a positive constant. Otherwise, if the separation only grows linearly, does not grow, or even decreases, the lyapunov exponent will be either extremely close to 0, exactly 0, or negative, all of which meaning the orbit is not chaotic. To sum it up:

- If $\lambda > 0$, this means the gap is growing exponentially, showing that the orbit is chaotic. The larger λ, the faster the two trajectories it is measuring are separating.
- If $\lambda \leq 0$, this means that the gap is either growing linearly or is bounded, showing that the motion is regular instead of chaotic. This is what we are expecting from the integrable system.

However a problem arises. Since the energy of our system is bounded, two orbits can only separate so far before they run out of room. Once the gap grows to the size of the possible region, it saturates, resulting in the exponential rate of separation disappearing. This causes the measurement of λ to be badly underestimated. To fix this we use a new method of renormalizing the separation periodically:

- First start a pair of orbits a tiny distance δ_0 apart and integrate both by a short interval T_{renorm}
- Measure the new separation, δ_k, and record $\ln(\delta_k/\delta_0)$
- Next, while keeping the velocities, reset the positions to δ_0. Repeat this process until the Lyapunov exponent converges to a stable value. This method will prevent the underestimation from the limited phase space available due to the limited energy.

Averaging the values recorded for each stretch of T_{renorm} gives a good approximation the lyapunov exponent:

$$\lambda \approx \frac{1}{NT_{renorm}} \sum_{k=1}^{N} \ln\left(\frac{\delta_k}{\delta_0}\right)$$

Results and Discussion

In the results and discussion section, the Poincaré sections of multiple systems with varying q values will be shown. Additionally, to explore one of the core concepts of chaos, sensitivity to initial conditions, a path separation vs time of an integrable system and a non-integrable system will be compared. Lastly, a MLE vs axis ratio graph will be plotted to observe the trend of how chaotic the system is as it progressively gets more squished due to

lower q values.

Parameters for Simulation

Below are the parameters used for the Poincaré sections:

- $v_c = 1.0$
- $R_c = 0.14$
- Energy, E = -0.337
- # of orbits = 374 (22 x_0 values, and 17 v_{x0} values for each x_0, creating a set of 374 initial conditions)
- rtol = 10^{-9}

Here are the parameters used for the Lyapunov exponent calculations:

- Initial separation, $\delta_0 = 10^{-3}$
- Renormalization interval, $T_{renorm} = 10$
- Total time, T = 5000
- 5 orbits used for Lyapunov exponent plots, taking the highest value as the MLE

System A (q = 1.0)

Below are the Poincaré section and the MLE convergence graph for System A, where the axis ratio is q = 1.0.

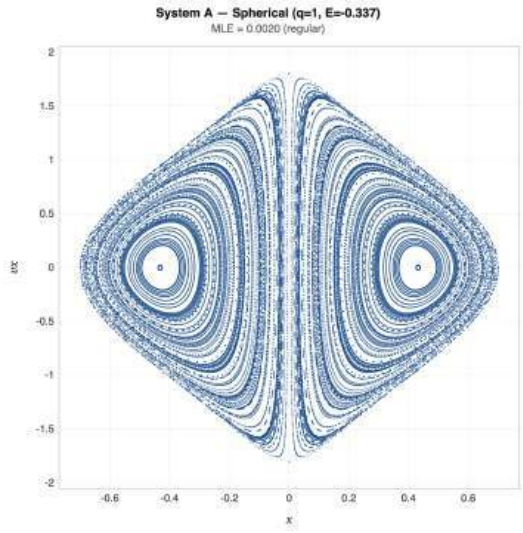


Fig 1: Poincaré section of System A (q=1.0)

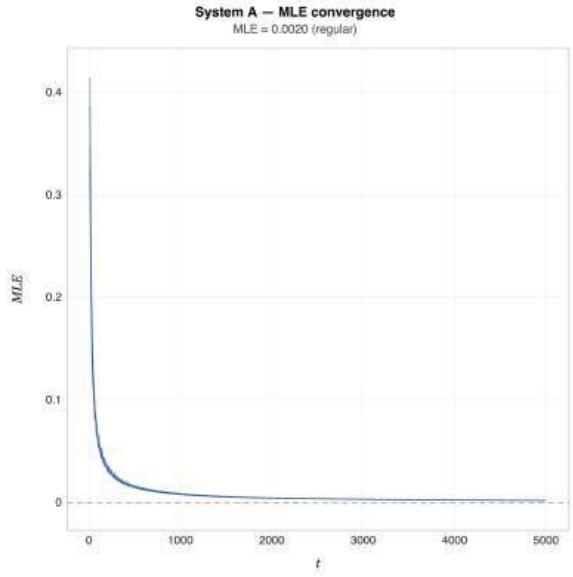


Fig 2: MLE convergence of System A (q=1.0)

The Poincaré section shows the expected enclosed loops around two symmetrical centers on either side of the section. This corresponds to many invariant tori in the phase space around which the trajectories of each orbit in the phase space wraps around. An important note for this is although the points may look scattered, it only arises from the orbits being close together and not enough time integrated for the longer loops around the outside to be completely covered, and not because the q = 1.0 system is chaotic. This Poincaré section completely matches what is expected of a fully integrable system.

Moving on to look at the graph of the MLE. 5 pairs of orbits were calculated for this graph. Since this is an integrable system, all of the Lyapunov exponent values quickly converged towards 0. While the MLE value of 0.0020 is not the exact 0 that was expected, this is not an issue since the renormalization method is only an approximation that approaches MLE = 0 as the total integration time, T, approaches infinity. Since the MLE is such a low value here, it matches the near 0 value that comes with an integrable system.

System B (q=0.6)

Below are the Poincaré section and the MLE convergence graph for System B, where the axis ratio is q = 0.6.

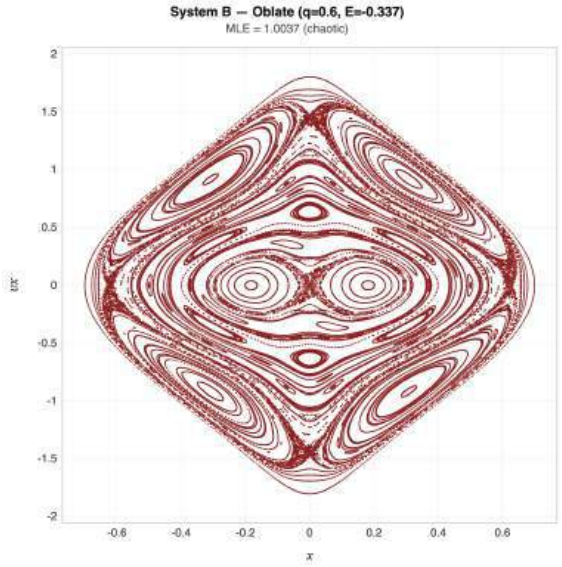


Fig 3: Poincaré section of System B (q=0.6)

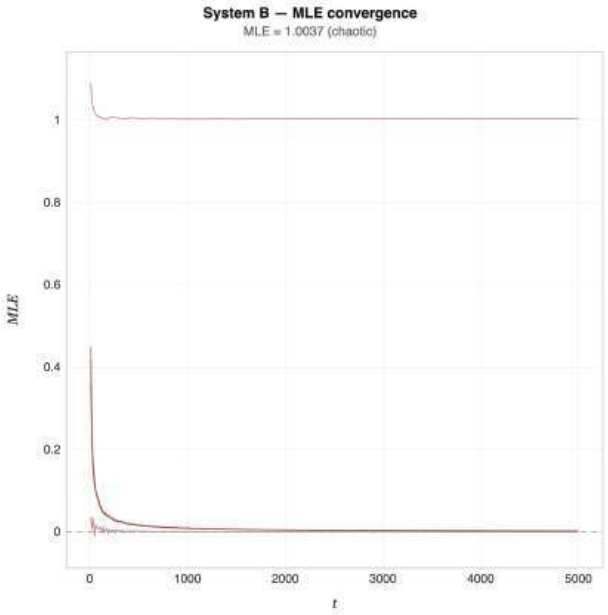


Fig 4: MLE convergence of System B (q=0.6)

In contrast to the clean and nested loops of System A, this Poincaré section is a mix of many separate regions of nested loops and chaotic regions in between with scattered points. As predicted by the KAM theory, some of the tori survived the loss of a conserved quantity, only becoming distorted instead of completely breaking. However, the chaos can still be clearly seen in the many tori that completely broke causing the previously quasiperiodic orbit to become chaotic.

The MLE for this system is 1.0037, which also matches expectations for a chaotic

system. In the graph, 4 out of the 5 lines are from the surviving quasiperiodic orbits, which converge to 0 in the graph shown, but the MLE comes from the one orbit pair that is chaotic near the top of the graph, quickly converging to a Lyapunov exponent of around 1 soon after the start of the simulation.

Path Separation

This is a graph of the natural log of the separation on the y axis vs time on the x axis. Firstly, an important note is that this graph does not use the renormalization method, since it does not really make sense to have a separation greater than the maximum length of the potential. The renormalization method was only used for the MLE calculations since without it there would be a significant underestimation.

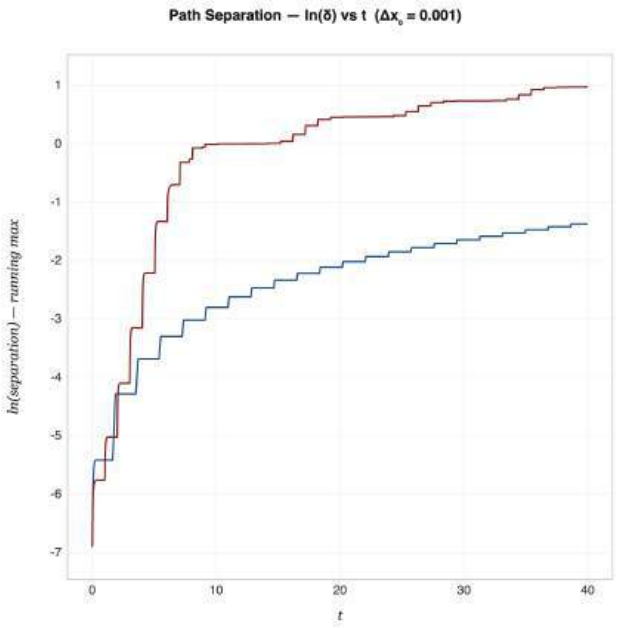


Fig 5: Natural log of separation vs time of Systems A and B

The blue line resembling a log function is the running max of the separation of an orbit in System A. A log function on the log scale corresponds to linear growth if both the axes are regular, matching the expectation of an integrable system where the separation grows linearly. The red line is the separation for System B. The line is straight up until approximately t = 8, when the separation saturates as it nears the maximum available space. Since a straight line on a log-scaled y axis corresponds to an exponential growth in a regular graph, this satisfies the expectations of the exponential separation of a chaotic orbit.

MLE vs Axis Ratio

Here we plot a MLE vs axis ratio graph to observe the trend in MLE as the axis ratio changes.

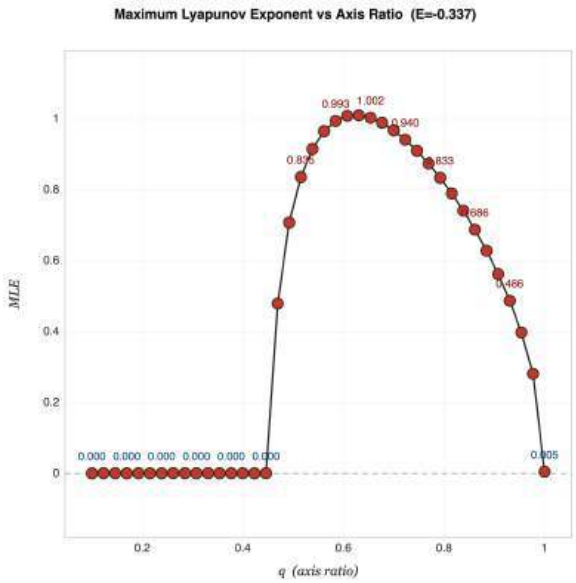


Fig 6: MLE vs axis ratio

Surprisingly, the plot is not monotonic. Instead of the expected continuous rise in MLE as the potential gets further squished by decreasing axis ratios, the MLE peaks at around q=0.6, and reaches 0 once again around q=0.45. This trend could result from the potential being squished too much to the point where it starts approximating a 1D system instead of 2D. Since the potential is squished along the y axis, the mass still orbits slowly in the x axis. However, since the potential is significantly thinner in the y direction compared to the x direction, the mass only oscillates in the y direction around the x axis while most of the motion happens in the x direction. Since energy is still conserved, this is enough to make a 1D system integrable. However, this system, though it has an MLE less than even the q=1.0 system, is not truly integrable, since it only approximates a 1D system instead of truly becoming one; this means that the systems where $q < 0.45$ are near-integrable but not truly, since there is still motion in the y direction, just not significant enough to break the regular motion.

To verify this trend, the Poincaré section and the MLE calculation are created to check qualitatively if the near-integrability is just a calculation error, after changing the axis ratio of System B, the oblate system, to q=0.3

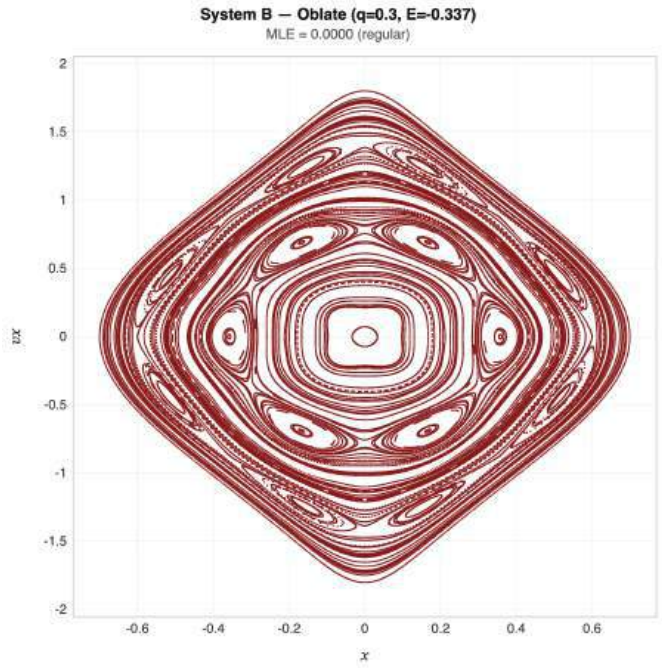


Fig 7: Poincaré section of System B (q=0.3)

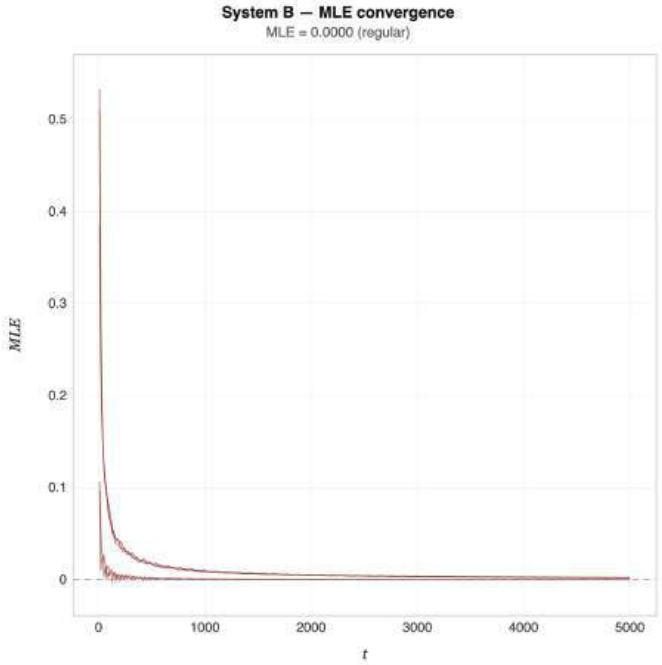


Fig 8: MLE convergence of System B (q=0.3)

While this Poincaré section is significantly more complicated than the one of the fully integrable system with q=1.0, it is indeed filled with loops like the extremely low MLE value suggests. This also supports the theory of fast y oscillations making the system approximate a 1D

potential: this Poincaré section has a lot more crossings in this y=0 section compared to System A, as can be seen in the the thicker, darker lines all around the shape in the section. This implies that the mass is quickly passing through the plane many times faster in the same amount of time compared to the truly integrable case.

Conclusion

In summary, integrability comes from conserved quantities, which originates from their respective symmetries within the system. This paper specifically studies a 2D logarithmic potential system, often used in galactic mechanics, as the axis ratio of the system squishes the potential. This squishing of the potential breaks the symmetry required to produce sufficient conserved quantities to keep the system integrable. The trend was successfully examined through the use of Poincaré sections as a qualitative measure and Lyapunov exponents as a quantitative measure.

While most of the results matched with expectations, the most surprising result of the paper comes from the MLE vs q of the system not being monotonic. Initially, the system was expected to be more chaotic as it continuously gets squished by lowering the axis ratio. However, the MLE value peaks and hits 0 once again around q=0.45. This phenomenon was attributed to the potential getting overly squished causing the motion of test particles to approximate that of a 1D system.

For future research, the reason behind why the MLE starts to flatten out at approximately q=0.45 can be further explored. This can be done by tracking the path of the test mass in physical space to verify if the flattened potential truly does approximate a 1D system. Another further possibility for exploration for the simulation would be to break the time translation symmetry as well, making energy no longer conserved. This can be done by rotating the potential with respect to time, causing the forces on the test mass to be different at the same position for different times.

AI Acknowledgement

AI was used for creating the simulation for the plots used in this paper. It was not used for any of the writing of the manuscript. The file for the simulation has been linked inside this anonymous repository below.
https://anonymous.4open.science/r/Research-Simulation-File-BDAD/poincare_explorer.html

Works Cited

Arnol′d, V. I., et al. Mathematical Methods of Classical Mechanics. Springer, 1989.

Cincotta, P. M., and C. Simó. “Simple Tools to Study Global Dynamics in Non-Axisymmetric Galactic Potentials – I.” Astronomy and Astrophysics Supplement Series, vol. 147, no. 2, EDP Sciences, Dec. 2000, pp. 205–28, https://doi.org/10.1051/aas:2000108. Accessed 25 July 2026.

J Bruce Brackenridge, et al. The Key to Newton’S Dynamics : The Kepler Problem and the Principia. University Of California Press, Cop, 1995.

Lorenz, Edward N. “Deterministic Nonperiodic Flow.” Journal of the Atmospheric Sciences, vol. 20, no. 2, Mar. 1963, pp. 130–41, https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2. Accessed 24 July 2026.

Lyapunov, A. M. General Problem of the Stability of Motion. CRC Press, 1992.

Noether, Emmy. “Invariant Variation Problems.” Transport Theory and Statistical Physics, vol. 1, no. 3, Jan. 1971, pp. 186–207, https://doi.org/10.1080/00411457108231446.

Oestreicher, Christian. “Chronobiology in Psychiatry.” Dialogues in Clinical Neuroscience, vol. 9, no. 3, Sept. 2007, https://doi.org/10.31887/dcns.2007.9.3.

Olivier Babelon, et al. Introduction to Classical Integrable Systems. Cambridge University Press, 2003.

Rothman, D. Lecture Notes for 12.006J/18.353J/2.050J, Nonlinear Dynamics: Chaos. 2022, https://ocw.mit.edu/courses/12-006j-nonlinear-dynamics-chaos-fall-2022/mit12_006jf22_lec15-16.pdf. Accessed 25 July 2026.

Williams, Garnett. Chaos Theory Tamed. Crc Press, 2019.

Wolf, Alan, et al. “Determining Lyapunov Exponents from a Time Series.” Physica D: Nonlinear Phenomena, vol. 16, no. 3, July 1985, pp. 285–317, https://doi.org/10.1016/0167-2789(85)90011-9.

The Impact of Limited English Proficiency on Healthcare Access Across Six Domains in the United States By Shruthi Shanker

Abstract

Limited English proficiency, or LEP, is a well known barrier to healthcare access in the United States, but most existing research studies it within a single clinical setting or a single health outcome. This paper asks a broader question: does LEP's relationship with healthcare access look the same across different kinds of care, or does it change from one domain to the next? Using a county level dataset built from the CDC/ATSDR Social Vulnerability Index, CDC PLACES small area health estimates, and CDC COVID-19 county vaccination records for all 3,144 United States counties, this study examines six outcomes: health insurance coverage, self reported cost barriers to care, mammography screening, colorectal cancer screening, annual checkups, and COVID-19 vaccine series completion. After controlling for county poverty rate, a county's LEP rate is positively associated with its uninsured rate (standardized $\beta = 0.31$, 95% CI [0.28, 0.34], $p < .001$) and even more strongly associated with residents reporting they could not see a doctor because of cost ($\beta = 0.54$, 95% CI [0.52, 0.57], $p < .001$). LEP's relationship with preventive screening is weaker and, in most cases, smaller than poverty's (mammography $\beta = -0.08$ versus poverty $\beta = -0.42$; colorectal screening $\beta = -0.26$ versus poverty $\beta = -0.55$; both LEP effects are significant at $p < .001$ despite their modest size). The most surprising finding is that COVID-19 vaccine completion is actually higher, not lower, in high LEP counties net of poverty ($\beta = 0.23$, 95% CI [0.20, 0.26], $p < .001$), a reversal that most likely reflects where high LEP counties tend to sit (dense, well resourced, immigrant gateway metro areas) rather than any real advantage that limited English proficiency itself confers. Taken together, these results suggest that policy responses to LEP related health disparities should be designed domain by domain rather than treating “the LEP barrier” as one uniform problem.

1. Introduction

As of 2015, an estimated 25.9 million people in the United States spoke English less than “very well,” the U.S. Census Bureau's threshold for limited English proficiency (Batalova and Zong). This population is not spread evenly across the country. It clusters in specific counties, largely along the southern border, in agricultural regions, and in large immigrant gateway metro areas. A substantial body of research links limited English proficiency to worse healthcare experiences and outcomes. Flores found that language barriers are tied to reduced access to care, lower patient satisfaction, and worse quality of care across a range of studies (229). Karliner et al., in a systematic review, found that professional interpreter use improves clinical care for patients with limited English proficiency, which implies that going without such services makes care worse (727). Fiscella et al. showed that language based disparities in care persist even among people who already have insurance, suggesting the barrier operates beyond simple enrollment (52), and DuBard and Gizlice found language associated gaps in specific preventive services among U.S. Hispanic adults that held even after accounting for insurance status.

What existing research rarely does is compare the size of the LEP barrier across different kinds of healthcare access within the same population and the same time period. Knowing that LEP matters for, say, interpreter dependent primary care visits does not tell us whether it matters just as much for enrolling in insurance, for cost related avoidance of care, for cancer screening uptake, or for vaccination. If the barrier is roughly uniform, a single policy lever, broader interpreter mandates for instance, might be expected to work everywhere. If the barrier is domain specific, effective interventions probably need to be domain specific too.

This study takes on that question descriptively, using publicly available, county level data for all 3,144 counties and county equivalents in the United States. It combines a language, poverty, and insurance dataset (the CDC/ATSDR Social Vulnerability Index) with two additional CDC data sources covering preventive care and vaccination, and asks where, across insurance coverage, cost barriers, three types of preventive care, and COVID-19 vaccination, the association between a county's LEP rate and its healthcare access outcome is largest, where it is smallest, and where, if anywhere, it flips sign entirely.

2. Data and Methods

2.1 Data Sources

Three public data sources were merged at the county level using Federal Information Processing Standards (FIPS) codes.

The CDC/ATSDR Social Vulnerability Index (SVI), 2022 release, provided four county level variables drawn from the 2018 to 2022 American Community Survey: the percentage of residents age 5 and older who speak English “less than well” (the LEP variable), the percentage of residents below 150% of the federal poverty line, the percentage of the civilian noninstitutionalized population without health insurance, and the percentage of households without a broadband internet subscription. The SVI methodology is described in Flanagan et al., who developed the index for emergency management and disaster response, though its underlying Census derived variables are widely reused in health disparities research.

CDC PLACES: Local Data for Better Health, County Data, 2025 release, provided four more variables: the age adjusted percentage of women ages 50 to 74 up to date on mammography, the age adjusted percentage of adults ages 45 to 75 up to date on colorectal cancer screening, the age adjusted percentage of adults who reported a routine checkup within the past year, and the age adjusted percentage of adults who reported being unable to see a doctor in the past year because of cost. PLACES estimates are themselves modeled small area estimates, produced by applying multilevel regression and poststratification to the Behavioral Risk Factor Surveillance System (BRFSS) rather than by tallying direct county level survey counts. This method and its validation against direct survey estimates are described in Zhang et al. (127), and the modeling step is discussed further in the Limitations section.

CDC COVID-19 Vaccinations in the United States, County provided the percentage of each county's total population with a completed primary COVID-19 vaccine series, using the last

widely reported snapshot date of May 10, 2023, after which county level reporting became noticeably less complete.

After merging on FIPS code, coverage ranged from 2,956 counties (for the annual checkup and cost barrier measures, which PLACES suppresses in some small population counties) to 3,144 or 3,143 counties for the remaining measures. No values were imputed. Every analysis below uses the counties with non missing data for the variables involved, and sample sizes are reported individually.

2.2 Analytic Approach

Because LEP and poverty could plausibly track each other (concentrated poverty and concentrated immigrant populations often coincide geographically, at least in popular perception), the first analytic step was to check that assumption directly. Across the 3,144 counties, LEP rate and poverty rate turn out to be nearly uncorrelated ($r = 0.06$, $p = .001$, $n = 3,144$). High LEP counties are not disproportionately high poverty counties in this data. That independence makes it possible to estimate LEP's association with each outcome net of poverty without the two variables mostly capturing the same underlying pattern, and it is shown directly in Figure 2.

For each of the six outcomes, a linear model of the form Outcome ~ LEP% + Poverty% was fit using ordinary least squares (OLS). Two versions of each LEP coefficient are reported throughout this paper, and they are not the same number, so it is worth spelling out the difference. The raw, unstandardized coefficient, denoted b, is in the outcome's original percentage point units. It says how many percentage points the outcome is predicted to change for each one percentage point increase in LEP rate. The standardized coefficient, denoted β, rescales both the predictor and the outcome into standard deviation units before fitting, so that effects measured on very different original scales (percentage points of insurance coverage, of screening uptake, of vaccination) become directly comparable. A β of 0.31 means that a one standard deviation increase in a county's LEP rate is associated with a 0.31 standard deviation change in the outcome, holding poverty constant. Because standardization is just a linear rescaling of the same underlying model, the raw coefficient b and the standardized coefficient β for a given predictor always carry the identical t statistic and p value. Only their units differ, not their statistical significance. Every regression result reported below includes the raw coefficient b, its standard error, a 95% confidence interval, and a p value, alongside the standardized β used for cross outcome comparison. These are reported in full in Table 1.

To see the shape of the LEP outcome relationship without distortion from the fact that most U.S. counties have very low LEP rates (median 0.7%, mean 1.6%) while a small number have very high rates (maximum 36.4%, in Kenedy County, Texas), counties were also grouped into ten equal sized bins, or deciles, by LEP rate, and the mean outcome value was computed within each decile. This decile binning approach can reveal nonlinearities, threshold effects in particular, that a single linear coefficient tends to obscure.

Finally, for two outcomes specifically, uninsured rate and cost barrier, each county's residual (its actual rate minus what the LEP and poverty model predicts for it) was computed and used to identify the counties whose rates are least explained by this model, in either direction.

All of these analyses are ecological. They describe relationships between county level aggregates, not relationships between individual people's English proficiency and their individual healthcare access, a distinction with a long history in the social sciences (Robinson 351) and one with a direct bearing on how the vaccination finding in Section 3.6 should be read. This distinction and its implications are discussed further in the Limitations section.

3. Results

3.1 Where Limited English Proficiency Is Concentrated

Figure 1 maps the county level LEP rate across the United States. Elevated LEP rates cluster in a recognizable set of places: counties along the Texas-Mexico border, California's Central Valley and southern border counties, South Florida (including Miami-Dade County), and a scattering of large immigrant gateway metro counties elsewhere in the country. The great majority of U.S. counties have very low LEP rates. Only 226 of 3,144 counties (7.2%) have an LEP rate above 5%, and the median county sits at 0.7%. This geographic concentration matters for how the rest of the results should be read, especially the vaccination finding in Section 3.6.

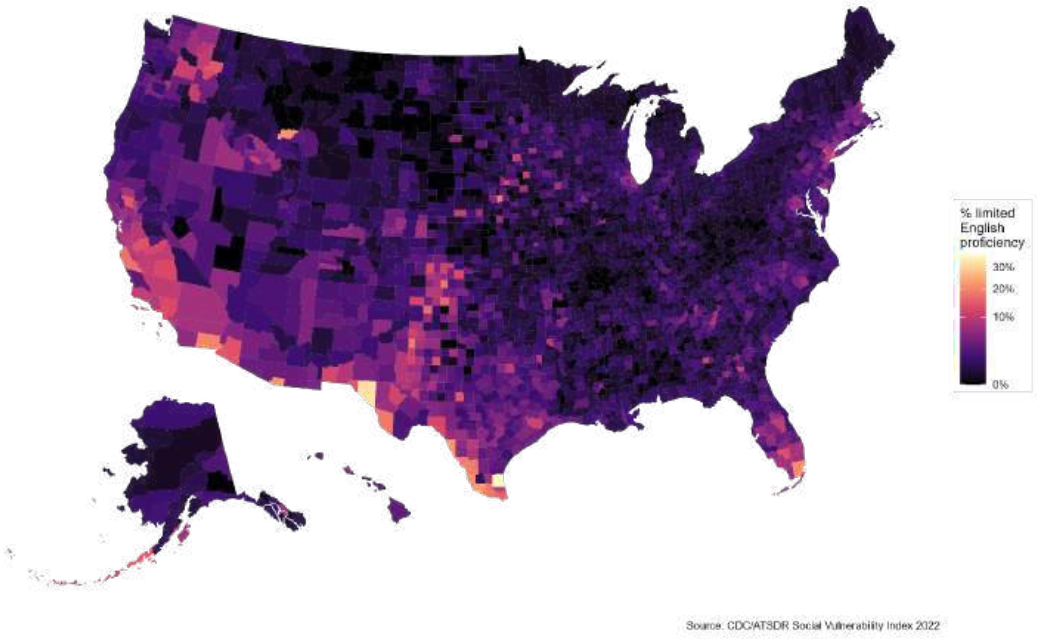


Figure 1. Geographic distribution of limited English proficiency (LEP) across the United States. Choropleth map of county-level LEP rate, defined as the percentage of residents age 5 and older who speak English less than "very well." The color scale uses a square-root transformation to accommodate the right-skewed distribution of LEP rates (median 0.7%, maximum 36.4%). n = 3,144 counties. Data source: CDC/ATSDR Social Vulnerability Index, 2022 release (2018-2022 American Community Survey).

3.2 Is LEP Just a Proxy for Poverty?

Before comparing LEP's effect to poverty's across six outcomes, it is worth confirming that the two are not simply measuring the same thing. Figure 2 plots every county's poverty rate against its LEP rate directly. The relationship is about as flat as a relationship can be while still being statistically detectable: r = 0.06 (p = .001, n = 3,144). A handful of high poverty counties also have unusually high LEP rates, and these are visible as the scattered points in the upper right of the figure, but the regression line barely rises across the full range of poverty values. In practical terms, knowing a county's poverty rate tells you almost nothing about its LEP rate, and vice versa. That independence is what allows the models below to attribute separate, largely non overlapping shares of each outcome to LEP and to poverty, rather than splitting one shared signal between two correlated variables.

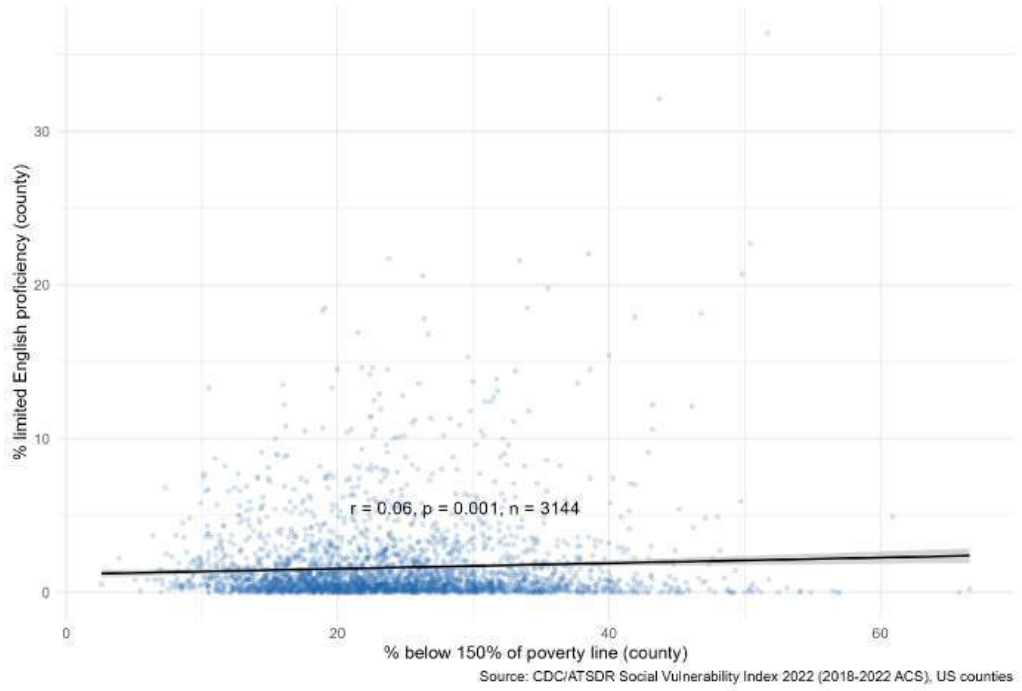


Figure 2. Limited English proficiency is only weakly related to poverty at the county level. Each point represents one county's LEP rate plotted against its poverty rate (percentage of residents below 150% of the federal poverty line). The black line is an ordinary-least-squares regression fit; the shaded band is its 95% confidence interval. Pearson r = 0.06 (p = .001), n = 3,144 counties. Data source: CDC/ATSDR Social Vulnerability Index, 2022 release.

3.3 An Overview Across All Six Domains

Before zooming in on any single outcome, Figure 3 lays out the shape of the LEP relationship for all six outcomes side by side, using the same decile binning approach introduced above. The uninsured rates and cost barrier panels (A and B) both rise toward the high LEP end of the range, most noticeably in the top two or three deciles. The three preventive care panels (C,

D, and E) stay comparatively flat across most of the LEP range, with mammography and annual checkup rates declining only modestly even in the highest LEP decile, and colorectal screening dipping mainly in the very highest decile. The vaccination panel (F) rises gradually and fairly steadily across the full range of deciles, most sharply between the ninth and tenth. Looking at all six panels together makes a point that will run through the rest of this paper: there is no single shape that describes how LEP relates to healthcare access in this data. Each domain has its own curve, and the next three figures unpack why.

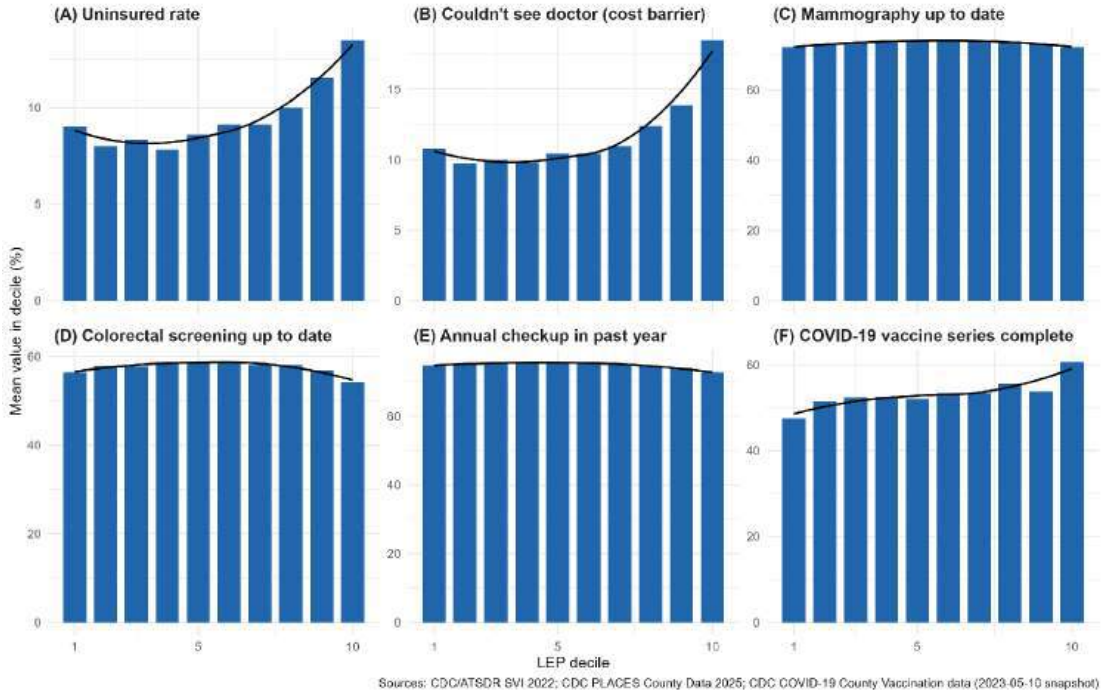


Figure 3. Mean value of six healthcare-access outcomes across limited-English-proficiency deciles. Counties were divided into ten equal-sized groups (deciles) ranked by LEP rate (decile 1, lowest LEP; decile 10, highest LEP), and the mean outcome value was computed within each decile. Bars show decile means; black curves show a loess-smoothed trend line (span = 1). (A) Uninsured rate, n = 3,144. (B) Percentage unable to see a doctor in the past year due to cost, n = 2,956. (C) Percentage of women ages 50-74 up to date on mammography, n = 3,143. (D) Percentage of adults ages 45-75 up to date on colorectal cancer screening, n = 3,143. (E) Percentage of adults reporting a routine checkup in the past year, n = 2,956. (F) Percentage of total population with a completed primary COVID-19 vaccine series as of May 10, 2023, n = 3,117. Data sources: CDC/ATSDR Social Vulnerability Index, 2022 release; CDC PLACES County Data, 2025 release; CDC COVID-19 Vaccinations in the United States, County.

3.4 Insurance and Cost Barriers Share a Threshold Shape

Figure 4 zooms in on the two outcomes with the clearest pattern in Figure 3: uninsured rate (panel A) and cost barrier (panel B), both plotted by LEP decile with standard errors and a loess trend line. Neither relationship is a smooth, straight line. Uninsured rates stay roughly flat, between about 8% and 9%, across the bottom seven deciles (LEP rates from 0% up to roughly 1.2%), then climb more steeply across the top three, reaching a mean of 13.5% in the tenth decile

(average LEP rate 7.8%). The cost barrier follows nearly the same shape: flat around 10% through the middle of the range, then rising to 18.4% in the tenth decile. In both cases, the outcome does not appear to respond to small increases in LEP rate until that rate crosses a rough threshold, after which the two move together much more visibly. This is consistent with the overall correlations (r = 0.34 for uninsured rate, r = 0.58 for cost barrier, across all counties), but the decile view shows that the relationship is concentrated among a minority of high LEP counties rather than spread evenly across the whole range.

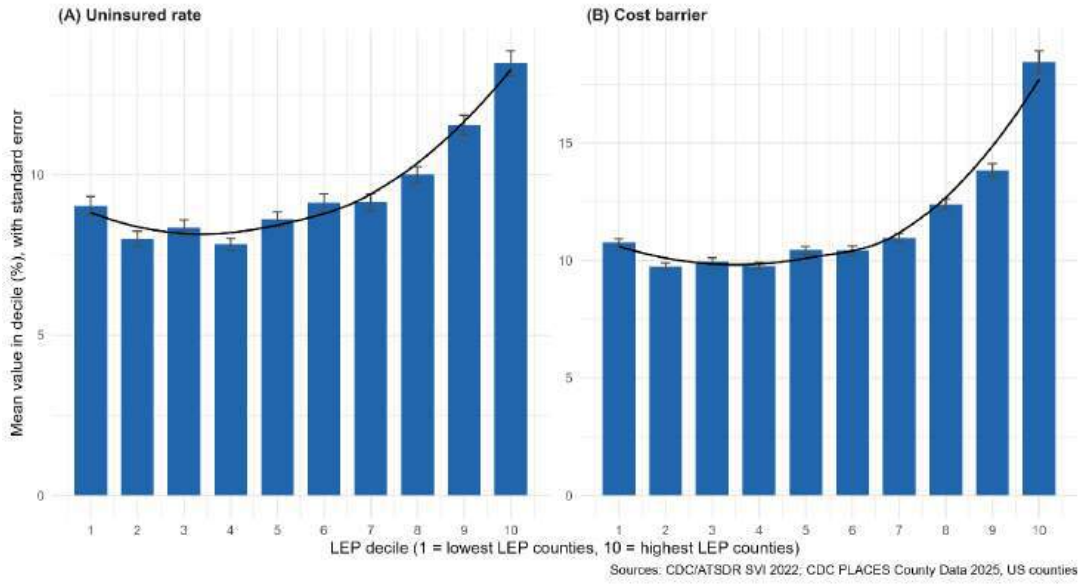


Figure 4. Uninsured rate and cost-barrier rate rise sharply only in the highest limited-English-proficiency deciles. Bars show the mean outcome value within each LEP decile; error bars show ± 1 standard error of the mean; the black curve is a loose-smoothed trend line (span = 1). (A) Uninsured rate, n = 3,144 counties. (B) Percentage unable to see a doctor in the past year due to cost, n = 2,956 counties. Data sources: CDC/ATSDR Social Vulnerability Index, 2022 release; CDC PLACES County Data, 2025 release.

3.5 Which Counties Defy the Model

Figure 5 identifies the sixteen counties (eight in each direction) whose actual uninsured rates deviate most from what the LEP and poverty model predicts (panels A and B), and does the same for cost barriers (panels C and D). Among counties with uninsured rates far higher than the model expects are LaGrange County, Indiana, and Holmes County, Ohio, both home to large Amish communities that, for religious reasons, largely decline commercial and public health insurance regardless of income or English proficiency, relying instead on community based mutual aid arrangements that predate, and were later exempted from, the Affordable Care Act's coverage requirement (Dundes). Also on this list are Buffalo County and Oglala Lakota County, South Dakota, both containing large Native American reservations where insurance coverage gaps trace to chronic underfunding of the Indian Health Service rather than to language or poverty as measured here (Khetpal et al.), along with several small counties (Scotland County,

North Carolina; Aleutians East Borough and Lake and Peninsula Borough, Alaska; Davis County) whose extreme values are also consistent with small population denominators.

Among counties with uninsured rates far lower than the model predicts are Bronx County, New York, and several small counties including Luna, Lee, and Leslie Counties, Imperial County, California, and Kenedy County, Texas, the same Kenedy County that has the highest LEP rate in the entire dataset (36.4%) yet one of the largest negative residuals, meaning its uninsured rate sits far below what its very high LEP rate and poverty rate alone would predict. Bronx County's position on this list fits with New York State's relatively expansive public insurance infrastructure, while several of the small counties on both sides of this list are plausibly driven substantially by small population volatility rather than by any systematic factor, a point discussed further in Section 5.

The cost barrier panels tell a more geographically concentrated story. Every county with a cost barrier rate far higher than expected (panel C) sits in South Texas: Jim Hogg, Zavala, Duval, La Salle, Jim Wells, Reagan, Dimmit, and Willacy Counties. These are rural, majority Hispanic border counties where a doctor visit can mean a long drive as much as a language barrier, and where the model's two predictors do not fully capture whatever combination of provider scarcity and rurality is pushing self reported cost barriers above what LEP and poverty alone would suggest. The counties with cost barrier rates far lower than expected (panel D) are almost the mirror image: dense, well resourced cities, including four New York City boroughs (Manhattan, Brooklyn, Queens, and the Bronx), San Francisco, Suffolk County, Massachusetts, and Clark County, Nevada (home to Las Vegas), plus Kenedy County again, this time on the low side for cost barrier just as it was for uninsured rate. Read together, these two sets of outlier counties suggest that cost barriers respond more to a county's healthcare infrastructure, urban or rural, than to LEP or poverty alone.

Taken as a whole, this figure's point is that both uninsured rate and cost barrier have multiple, distinct root causes (religious objection to insurance, the specific financing structure of tribal health coverage, rural provider scarcity along the border, small county statistical noise) that a two variable model of LEP and poverty cannot fully capture, and that a language access intervention alone would not be expected to move the needle in several of these specific places.

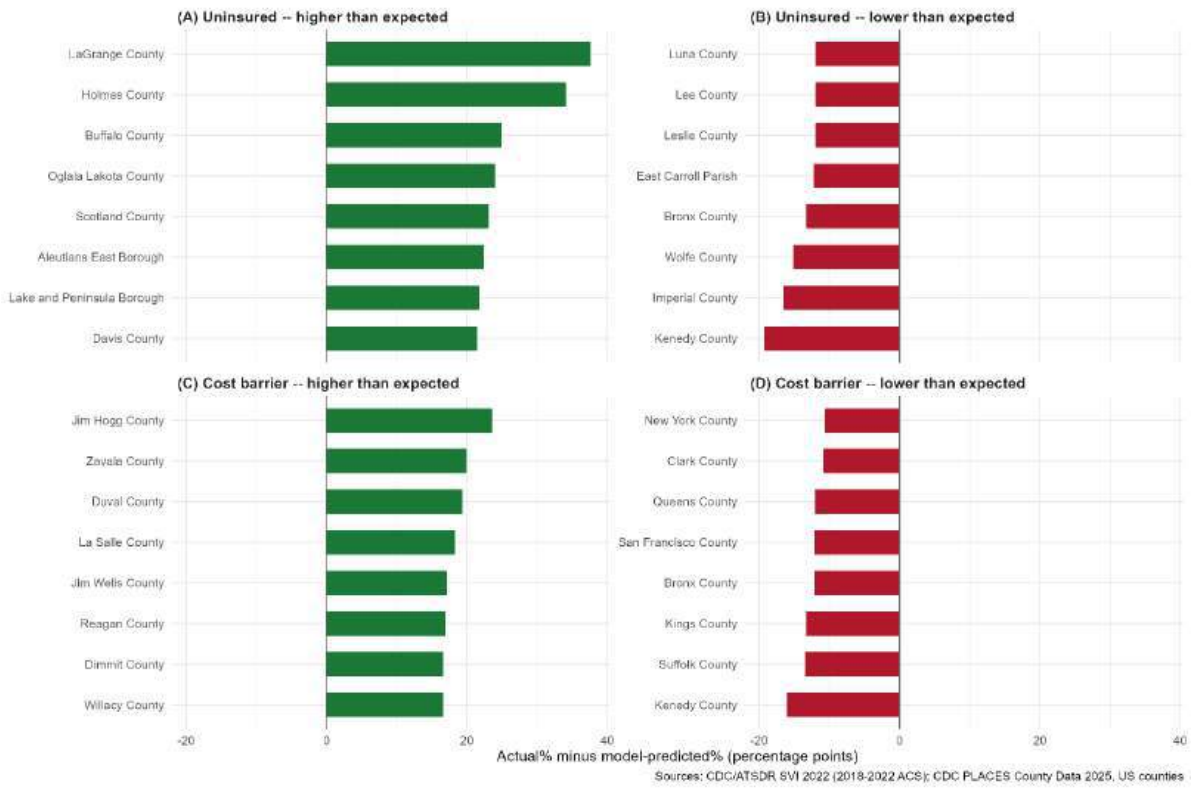


Figure 5. Counties whose uninsured rate or cost-barrier rate deviate most from the values predicted by a linear model of LEP and poverty. For each outcome, an ordinary-least-squares model of the form Outcome ~ LEP% + Poverty% was fit, and each county's residual (actual value minus model-predicted value) was computed. Bars show the eight counties with the largest positive residuals and the eight with the largest negative residuals for each outcome. (A) Uninsured rate, higher than predicted. (B) Uninsured rate, lower than predicted. (C) Cost barrier, higher than predicted. (D) Cost barrier, lower than predicted. Data sources: CDC/ATSDR Social Vulnerability Index, 2022 release; CDC PLACES County Data, 2025 release.

3.6 LEP Versus Poverty, All Six Domains at Once

Figure 6 is this study's central result. It plots the standardized coefficient for LEP (panel A) and for poverty (panel B), net of each other, across all six outcomes, with the same outcome order and 95% confidence intervals in both panels so the two can be compared row by row. The pattern is not uniform, and it does not even point in a single direction. LEP's largest association is with the cost barrier outcome, the share of residents who report being unable to see a doctor due to cost, at $\beta = 0.54$ (95% CI [0.52, 0.57], $p < .001$), which is larger than poverty's own standardized effect on that same outcome ($\beta = 0.48$). The uninsured rate association is smaller but still substantial ($\beta = 0.31$, 95% CI [0.28, 0.34], $p < .001$, versus poverty's $\beta = 0.40$). For the three preventive care outcomes the pattern changes. LEP's association with colorectal cancer screening ($\beta = -0.26$, 95% CI [-0.29, -0.23], $p < .001$) and annual checkups ($\beta = -0.24$, 95% CI [-0.28, -0.21], $p < .001$) is negative but noticeably smaller in size than poverty's association with the same outcomes ($\beta = -0.55$ and $\beta = 0.18$, respectively), and LEP's association with mammography screening is close to zero ($\beta = -0.08$, 95% CI [-0.11, -0.05], $p < .001$) next to

poverty's much larger negative effect (β = -0.42), a reminder that a small, tightly estimated effect is not the same thing as a practically meaningful one. For COVID-19 vaccine completion, the direction flips entirely: LEP's standardized effect is positive (β = 0.23, 95% CI [0.20, 0.26], p < .001) while poverty's is negative (β = -0.24), the only outcome among the six where LEP and poverty point in opposite directions. All six LEP coefficients are statistically significant at p < .001, which, with a sample of roughly 3,000 counties, is expected even for small effects. It is the confidence intervals, not the p values alone, that separate the practically large effects (cost barrier, uninsured rate) from the practically small ones (mammography).

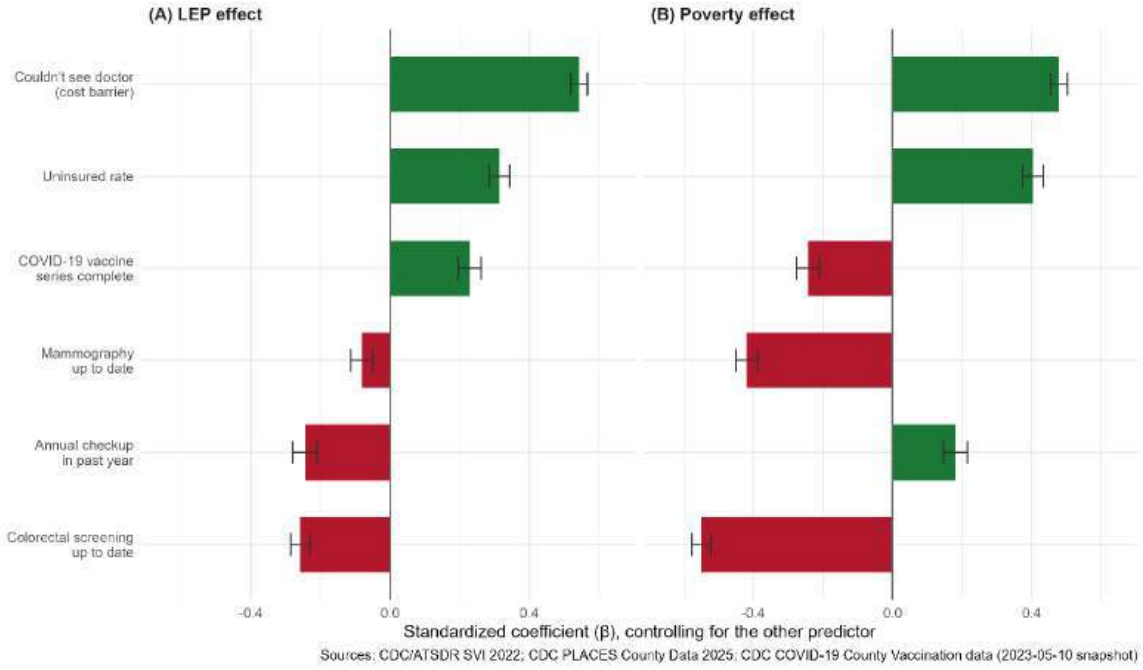


Figure 6. Standardized effect of limited English proficiency and of poverty on six healthcare-access outcomes. For each outcome, an ordinary-least-squares model of the form Outcome ~ LEP% + Poverty% was fit, and both predictors were standardized (rescaled to standard-deviation units) to allow direct comparison across outcomes measured on different scales. Bars show the standardized regression coefficient (β) for each predictor, net of the other; error bars show 95% confidence intervals. Bar color indicates the sign of the coefficient (dark green, positive; dark red, negative). Outcomes are ordered identically in both panels, ranked by the size of the LEP effect. (A) LEP effect. (B) Poverty effect. Data sources: CDC/ATSDR Social Vulnerability Index, 2022 release; CDC PLACES County Data, 2025 release; CDC COVID-19 Vaccinations in the United States, County, snapshot date May 10, 2023.

Table 1 reports the full model detail behind Figure 6 for every outcome: sample size, the raw (unstandardized) LEP coefficient b, its standard error, 95% confidence interval, and p value, alongside the standardized β for both LEP and poverty.

Outcome	n	b (LEP)	SE	95% CI (LEP)	p	β (LEP)	β (Poverty)
Uninsured rate	3,144	0.61	0.029	[0.28, 0.34]	<.001	0.31	0.40
Couldn't see doctor (cost)	2,956	0.96	0.021	[0.52, 0.57]	<.001	0.54	0.48
Mammography up to date	3,143	-0.13	0.026	[-0.11, -0.05]	<.001	-0.08	-0.42
Colorectal screening up to date	3,143	-0.43	0.023	[-0.29, -0.23]	<.001	-0.26	-0.55
Annual checkup in past year	2,956	-0.32	0.023	[-0.28, -0.21]	<.001	-0.24	0.18
COVID-19 vaccine series complete	3,117	1.12	0.083	[0.20, 0.26]	<.001	0.23	-0.24

Table 1. Regression results for Outcome ~ LEP% + Poverty% across six healthcare-access outcomes. b is the raw (unstandardized) LEP coefficient in outcome percentage points per percentage point of LEP; SE and the 95% CI apply to b; β is the standardized coefficient for LEP and poverty respectively, net of each other. All six LEP coefficients are significant at $p < .001$.

4. Discussion

4.1 LEP's Barrier Is Largest Where It Meets Cost and Enrollment

The two outcomes with the largest standardized LEP effects, cost related access barriers and insurance coverage, both involve navigating an administrative system: enrolling in a health plan, understanding what a plan covers, or deciding whether a given visit is affordable before or during a healthcare encounter. This lines up with Fiscella et al.'s finding that language based disparities in care persist even among people who already have insurance (52), which suggests the mechanism includes but is not limited to enrollment itself. It also fits the broader logic of Karliner et al.'s review, which found that professional interpretation improves clinical care specifically in encounter based, communication dependent settings (727), exactly the kind of setting where a cost barrier judgment or an insurance enrollment decision gets made.

4.2 LEP's Barrier Is Smaller, and Secondary to Poverty, for Preventive Screening

For mammography, colorectal cancer screening, and annual checkups, poverty is consistently the larger of the two standardized effects, and for mammography specifically, LEP's association is close to zero. This broadly matches DuBard and Gizlice, who found that among U.S. Hispanic adults, Spanish speakers lagged behind English speakers on some preventive service measures but not others, and that the size of the language associated gap varied by which specific service was examined, a pattern that mirrors this study's finding that LEP's effect size is not constant across preventive care outcomes. One plausible explanation is that these outcomes are shaped more by supply side factors, the physical availability of screening facilities and primary care capacity in a county, which track poverty and rurality more directly than they track language. Another possibility is that some of these screening programs, mammography in particular, which has been the subject of sustained public health outreach for decades, have built

up multilingual outreach infrastructure that has narrowed the specifically language driven gap even where the poverty driven gap remains large. This study's data cannot distinguish between these explanations, which is a natural direction for future research using individual level data.

4.3 The Vaccination Reversal Looks Like Geography, Not a Language-Access Success Story

The finding that COVID-19 vaccine completion rises with LEP, net of poverty, is the most counterintuitive result in this study, and it should not be read as evidence that limited English proficiency somehow helps vaccine uptake. Individual level evidence points the other way. Quadri et al., studying over 850,000 patients in a single health system, found that patients with limited English proficiency or a non English language preference experienced delayed vaccine uptake and higher rates of COVID-19 hospitalization and death than English speaking patients. That an individual level study finds LEP associated with worse vaccination outcomes, while this county level study finds the opposite association, is itself instructive. It is a textbook illustration of the ecological correlation problem Robinson described more than seventy years ago, in which a relationship observed across aggregated geographic units can run opposite to the relationship observed across individuals within those units (351). The more plausible explanation for this study's county level finding lies in where high LEP counties are located rather than in any effect of language proficiency itself. As Figure 1 shows, the counties with the highest LEP rates are disproportionately large, dense, immigrant gateway metro counties, Miami-Dade being one example, that also tend to have above average healthcare infrastructure, hospital density, and public health capacity, and that were often specific targets of well funded, multilingual COVID-19 vaccination campaigns during 2021 and 2022. In other words, LEP rate in this county level analysis is very likely functioning partly as a proxy for “large coastal or border metro area” rather than purely as a measure of a language barrier, and vaccination is the one domain in this study where that proxy relationship visibly overwhelms whatever negative effect the language barrier itself may have, an effect that the Quadri et al. individual level study suggests is real, just invisible, or even reversed, once the analysis moves to the county level.

4.4 Uninsurance and Cost Barriers Have Multiple, Distinct Root Causes

The outlier analysis in Figure 5 reinforces a point that is easy to lose in an aggregate model: the same national statistic, whether uninsured rate or cost barrier, is produced by communities with very different underlying stories. Religious communities that decline commercial insurance as a matter of practice (Dundes), tribal communities whose coverage gaps trace to the specific financing structure of the Indian Health Service (Khetpal et al.), a cluster of rural South Texas border counties where cost barriers likely reflect provider scarcity as much as language or poverty, and dense urban counties with expansive public safety net infrastructure are all visible in this single figure, and none of their positions on the list is primarily explained by English proficiency. This suggests that a language access intervention, however well designed, would do little to move the uninsured rate or the cost barrier rate in several of the specific counties that most drive the national extremes on either measure.

5. Limitations

This study has several limitations that bear on how its findings should be used.

First and most importantly, all analyses are ecological. They describe relationships between county level rates, not relationships between individual people's English proficiency and their individual healthcare access. It is possible, for instance, that within a high LEP county it is specifically the English proficient residents who are driving a favorable outcome, or the reverse. County level correlations cannot rule this out, a limitation formally described by Robinson as the gap between ecological and individual correlations (351) and often called the ecological fallacy. Section 4.3's vaccination finding is a direct illustration of why this matters: an individual level study of the same pandemic (Quadri et al.) found LEP associated with worse, not better, vaccination outcomes. Individual level data, such as linked survey and claims data, would be needed to test whether this study's other patterns hold at the person level as well.

Second, the CDC PLACES measures used for preventive care and cost barriers are themselves modeled small area estimates rather than direct county level survey counts. They are produced by applying multilevel regression and poststratification to the state representative Behavioral Risk Factor Surveillance System (Zhang et al. 127), and that modeling step introduces its own uncertainty, which is not reflected in the standard county level point estimates used here.

Third, this study controls for only one potential confounder, county poverty rate. Other plausible confounders, including rurality, the size and composition of a county's immigrant population specifically (as opposed to LEP status generally, which includes some U.S. born residents), state level Medicaid expansion status, and local healthcare workforce density, are not included, and any of them could partly explain the patterns reported here.

Fourth, the COVID-19 vaccination measure reflects a single outcome, a specific vaccine during a specific pandemic response with unusually intensive multilingual outreach, at a single snapshot in time (May 10, 2023). It should not be assumed to generalize to routine adult or childhood vaccination coverage outside the pandemic context.

Fifth, several of the counties identified in the Figure 5 outlier analysis have small total populations, which can produce extreme percentage values from small numerators and denominators alone, independent of any systematic factor. Kenedy County, Texas, which has both the highest LEP rate and one of the most favorable residuals for both uninsured rate and cost barrier in the dataset, had a 2020 Census population of 350 people, and its position at these extremes should be read with that in mind.

Finally, LEP itself is a self reported Census Bureau category (English spoken “less than very well”) rather than a direct, validated assessment of English proficiency, and it does not distinguish between recent arrivals still acquiring English and longtime residents who may have chosen not to fully assimilate linguistically for reasons unrelated to how long they have lived in the United States.

6. Conclusion

Limited English proficiency's relationship with healthcare access in the United States is not one story but several. Using county level data spanning insurance coverage, cost related access barriers, three types of preventive screening, and COVID-19 vaccination, this study finds that LEP's standardized association with the outcome is largest for cost and insurance related barriers, smaller and secondary to poverty for preventive screening, and reversed for vaccination completion, a reversal most plausibly explained by the geographic concentration of high LEP counties in well resourced metro areas rather than by any advantage that limited English proficiency confers, and one that individual level research (Quadri et al.) suggests may run the other way once the analysis moves below the county level. Two practical implications follow. Interventions aimed at reducing LEP related health disparities are likely to have the largest measurable effect if targeted at insurance enrollment and cost navigation assistance specifically, where this study finds the largest independent LEP effect. And researchers or policymakers who treat “the LEP barrier” as a single, uniformly sized problem risk drawing conclusions from whichever domain happens to have been studied, when the underlying reality, as this analysis of six domains within the same population and time period shows, is considerably more uneven.

Works Cited

Agency for Toxic Substances and Disease Registry. "Social Vulnerability Index." Centers for Disease Control and Prevention, 22 July 2024, www.atsdr.cdc.gov/place-health/php/svi/index.html. Accessed 26 July 2026.

Batalova, Jeanne, and Jie Zong. "Language Diversity and English Proficiency in the United States." Migration Policy Institute, 11 Nov. 2016, www.migrationpolicy.org/journal/spotlight/language-diversity-and-english-proficiency-united-states. Accessed 26 July 2026.

Centers for Disease Control and Prevention. "COVID-19 Vaccinations in the United States, County." Data.CDC.gov, data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh. Accessed 26 July 2026.

Centers for Disease Control and Prevention. "PLACES: Local Data for Better Health, County Data, 2025 Release." Data.CDC.gov, data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb. Accessed 26 July 2026.

DuBard, C. Annette, and Ziya Gizlice. "Language Spoken and Differences in Health Status, Access to Care, and Receipt of Preventive Services among US Hispanics." American Journal of Public Health, vol. 98, no. 11, 2008, pp. 2021-28.

Dundes, Lauren. "Sharing the Load: Amish Healthcare Financing." Healthcare, vol. 4, no. 4, 2016, article 92.

Fiscella, Kevin, et al. "Disparities in Health Care by Race, Ethnicity, and Language among the Insured: Findings from a National Sample." Medical Care, vol. 40, no. 1, 2002, pp. 52-59.

Flanagan, Barry E., et al. "A Social Vulnerability Index for Disaster Management." Journal of Homeland Security and Emergency Management, vol. 8, no. 1, 2011, article 3.

Flores, Glenn. "Language Barriers to Health Care in the United States." New England Journal of Medicine, vol. 355, no. 3, 2006, pp. 229-31.

Karliner, Leah S., et al. "Do Professional Interpreters Improve Clinical Care for Patients with Limited English Proficiency? A Systematic Review of Literature." Health Services Research, vol. 42, no. 2, 2007, pp. 727-54.

Khetpal, Vishal, et al. "A Federal Indian Health Insurance Plan: Fulfilling a Solemn Obligation to American Indians and Alaska Natives in the United States." Preventive Medicine Reports, vol. 25, 2021, article 101669.

Quadri, Nasreen S., et al. "Evaluation of Preferred Language and Timing of COVID-19 Vaccine Uptake and Disease Outcomes." JAMA Network Open, vol. 6, no. 4, 2023, article e237877.

Robinson, W. S. "Ecological Correlations and the Behavior of Individuals." *American* Sociological Review, vol. 15, no. 3, 1950, pp. 351-57.

Zhang, Xingyou, et al. "Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators from the Behavioral Risk Factor Surveillance System." American Journal of Epidemiology, vol. 182, no. 2, 2015, pp. 127-37.

Research Shows A Strong Inverse Correlation Between Religiosity and Per-capita Spending on Education. Does One Cause the Other? By Seoyeon Lee

If asked what education or knowledge is, many will likely think of "information", "credentials", "standardization", and possibly "ladder/success". Notably, concepts like the pursuit of ultimate truth, ethical commitment, or spiritual development are absent from this list. These are common perceptions. But are these the best definitions? Can we expand the boundaries of education and knowledge to include other dimensions of human development?

Before explaining the relationship itself, it is necessary to consider the underlying assumptions. Per-capita spending on education is often a reflection of a society's commitment to education. However, as suggested above, the current understanding of education may overlook certain dimensions of education, particularly those emphasized under religious frameworks. If society's commitment to education were defined to include those religious definitions of education, the relationship could be distinctly different. Hence, per-capita spending on education itself has important limitations.

To further contextualize the inverse correlation, an analysis of how education is perceived and how per-capita spending on education is measured is necessary. Through questioning the limited validity of per-capita spending on education, this essay argues that the inverse correlation between religiosity and per-capita spending on education cannot be comprehended straightforwardly as causation.

Current understandings of education are historically shaped. Although countries have various cultures and histories, modern nations face related challenges in economic competition, industrialization, and administration. As industrial economies require literate and technically skilled workers, modern nations depend on standardized systems for training, selection, and administration. To address these demands, as sociologist John W. Meyer and his colleagues suggest, modern schooling increasingly relies on mass, standardized, and measurable institutional structures, since these arrangements allow comparable and administratively useful forms of knowledge (Meyer et al., 1977).

Historical changes in the British education system illustrate this contingency. As education historian Sheldon Rothblatt (1976) describes, education in eighteenth- and nineteenth-century Britain was widely aimed at moral cultivation and civilization, as demonstrated in the importance of Latin, Greek, classics, and theology within school and university curricula. Consequently, standards of educational excellence were closely related to theological knowledge, classical learning, and the cultivation of moral judgement. However, industrialization transformed common understandings of educational aims. As scientific innovation, technical expertise, and economic productivity became increasingly central to social progress, educational priorities shifted toward science, mathematics, modern languages, and administration. Accordingly, the criteria for educational evaluation came to emphasize scientific competence, practical knowledge, and formal assessment.

As such, modern institutional forms became widespread and associated with social and

economic advancement, constructing public understanding of what education should look like. Meyer identifies this process as global convergence of modern schooling, noticing that mass schooling, science education, standardized curricula, and national examinations have become common worldwide (Meyer et al., 1977); sociologists Francisco O. Ramirez and John Boli further note that such characteristics of modern schooling have become closely associated with progress, development, and modernity (Ramirez & Boli, 1987).

Accordingly, current understandings of education privilege particular educational aims and forms of knowledge. The indicators used to evaluate educational success should also be understood as historically shaped, since they embed assumptions about what education should achieve. By choosing to measure certain outcomes rather than others, educational systems prioritize particular forms of knowledge and skill. As sociologist Jerry Z. Muller suggests, what societies choose to measure reveals what they value. Accordingly, educational aims associated with progress, development, and modernity are reflected through current educational indicators (Muller, 2018).

Current indicators in assessing educational success, such as literacy rates, standardized assessments, and per-capita educational spending, are designed to evaluate outcomes including technical competence, academic attainment, or economic productivity.

Interestingly, many educational aims emphasized within religious traditions are not commonly captured by current educational indicators. Although religious education varies across traditions, it often seeks not only to transmit religious knowledge, but also to cultivate moral character, shape worldviews such as on human flourishing, and guide individuals toward ultimate truths. Within such frameworks, education is frequently understood as a process of personal formation, having a different emphasis compared to modern schooling, by highlighting who people become. Thus, they differ in their relative emphasis and understanding of educational success.

Moreover, aims such as moral formation, spiritual development, and the search for meaning are more difficult to capture through indicators that are measurable, standardized, and comparable across populations. A student's compassion, moral judgement, or sense of purpose may be central to religious traditions, though such qualities rarely appear in mainstream educational indicators.

Specifically, dominant indicators of educational success tend to privilege educational outcomes that are measurable, standardized, and comparable, while making other educational aims less visible within public evaluations of education. The limited visibility of religious educational aims cannot be explained solely by the development of modern schooling; it also demonstrates the practical demands of modern governance, which relies on indicators that are standardized. For instance, modern schooling operates on large scales that require governments and international organizations to compare performance across institutions, to further allocate resources and evaluate policy outcomes. Such requirements depend on indicators to measure consistently in a standardized method. As historian and sociologist Theodore M. Porter argues, modern institutions rely on quantification because numerical indicators appear objective,

comparable, and are easier for evaluation (Porter, 1995). Although initiatives to assess character development, faith formation, and human flourishing do exist, such indicators remain relatively marginal and difficult to translate within mainstream educational evaluation frameworks. In contrast, modern educational systems, as mentioned, with more emphasis on how much one knows, are easier to quantify.

Educational indicators operate as a mirror and as a reinforcer. Indicators assess educational aims while shaping the public perception of educational success. Particularly, in terms of reinforcement, when success becomes linked to the values of modern schooling, educational institutions are incentivized to enhance their performance according to these criteria. As a result, educational institutions increasingly adapt their priorities to align with existing indicators (Espeland & Sauder, 2007). For example, university rankings are found to influence admissions policies and resource allocation. Further encouraging institutions to invest in areas that improve ranking performance, such as research output and selectivity. Similar dynamics can be observed at the government level, including how educational expenditure targets and international benchmarks increasingly become prioritized in policy priorities and resource allocation. Thus, indicators do not merely evaluate educational systems but also shape them. Sociologists Wendy Nelson Espeland and Michael Sauder (2007) describe this phenomenon as the reactivity of measures: once indicators become established, they affect institutional behavior and redefine success according to what is measured. Schools, governments, and international organizations similarly orient their practices toward improving measurable performance, reinforcing a particular understanding of educational success while making alternative educational aims less visible.

Accordingly, educational indicators should be understood not as neutral measures, since they have limitations as a product of historically specific assumptions about the aims of education. By privileging certain educational aims over others, they shape public understandings of educational success, influence the allocation of educational resources, and determine which forms of education are more likely to receive recognition and investment. Consequently, educational aims that are less associated with measurable indicators may become less visible within the dominant definitions of educational success.

This mechanism could explain the observed inverse correlation between religiosity and per-capita spending on education. Educational indicators assess educational success, implying what counts as education and which aims are publicly and educationally valued. As modern systems increasingly value aims associated with modern schooling, those of religious traditions become less visible in mainstream. In consequence, educational indicators can impact public understanding and patterns of resource allocation, shaping which forms of education receive greater recognition and investment. Thereby, the observed inverse correlation may display differences in educational priorities rather than a straightforward relationship between religiosity and per-capita spending on education. Such limited validity makes it necessary to examine further how per-capita spending on education is defined, measured, and compared across societies.

If educational indicators hold assumptions about what education should achieve, then educational spending itself requires closer examination. Per-capita spending on education is often treated as a neutral measure of educational commitment. However, such statistics do not merely capture how much societies value education; they also suggest investment in those institutions and programs recognized as educational within nationally defined educational frameworks. This, in turn, may impact which forms of education are publicly funded and presented in expenditure statistics.

In particular, international organizations such as UNESCO facilitate cross-national comparisons by aggregating data reported by national governments. However, these statistics depend upon how individual nations themselves define educational institutions and classify educational expenditure (UNESCO, 2024). Educational spending is measured by national definitions of education.

Differences in how societies define education complicate cross-national comparisons of education spending, as the same indicator may encompass fundamentally different forms of educational activity. For instance, Iran displays how educational activities may be classified differently across societies. In a nation with high religiosity and a theocratic system, as noted by sociologist Said Amir Arjomand, religious educational institutions receive substantial public funding, including Hawza seminaries and organizations associated with Qom, the country's principal centre of Shi'a scholarship (Arjomand, 1989). These institutions occupy an important place within Iran's educational landscape and receive support through state budget allocations. Whether and how such expenditures are incorporated into international education statistics, however, depends upon national reporting practices and the classifications adopted by organizations such as UNESCO.

In contrast, countries such as China largely organize public education through the Ministry of Education, where religious education exists outside the mainstream national educational framework. These differences suggest that educational spending cannot be interpreted independently of the institutional definitions through which education is recognized and funded. Furthermore, because educational indicators affect educational priorities and resource allocation, educational forms that align with dominant indicators of success are more likely to receive public recognition and investment. The same expenditure indicator may consequently capture fundamentally different educational realities across societies. Therefore, the observed inverse correlation between religiosity and educational spending may imply differences in how education is defined, measured, and funded across societies, complicating attempts to infer a straightforward, causal relationship between religiosity and per-capita spending on education. As a result, per-capita spending on education indicates investment only in forms of education recognized within national education systems.

In conclusion, the inverse correlation between religiosity and per-capita educational spending should not be interpreted as evidence that religious societies value education less or that education necessarily undermines religiosity. Rather, this argument cannot be understood independently of historically specific assumptions about how education is perceived and how

educational success should be measured. Hence, even though educational spending is a widely used indicator, understanding what it measures and what it overlooks is vital before drawing overall conclusions from it. Therefore, a straightforward causation cannot be proved by the observed correlation, since per-capita spending on education has limited validity as a measurement itself.

The argument presented aims to question the assumptions embedded within educational indicators, suggesting that such indicators do not fully capture the meaning of education and would be insufficient for establishing a direct causal relationship. It does not explain the changes in religious belief or the decline of religiosity itself.

Thus, the question is not whether religiosity and per-capita spending on education are inherently opposed. By reconsidering the assumption behind the definition of education, it invites a broader understanding of our current world and collective flourishing–one that highlights the acquisition of knowledge and skills, but also the cultivation of meaning, human formation, and human flourishing.

Works Cited

Arjomand, Said Amir. The Turban for the Crown : the Islamic Revolution in Iran. E-book ed., Oxford University Press, 1989.

Espeland, Wendy Nelson, and Michael Sauder. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." American Journal of Sociology, vol. 113, no. 1, 2007, pp. 1-40, https://doi.org/10.1086/522805.

Meyer, John W., et al. "The World Educational Revolution, 1950-1970." Sociology of Education, vol. 50, no. 4, 1977, pp. 242-58. JSTOR, https://doi.org/10.2307/2112498. Accessed 16 June 2026.

Muller, Jerry Z. The Tyranny of Metrics. Princeton University Press, 2018.

Porter, Theodore M. Trust in Numbers : the Pursuit of Objectivity in Science and Public Life. Princeton University Press, 1995.

Ramirez, Francisco O., and John Boli. "The Political Construction of Mass Schooling: European Origins and Worldwide Institutionalization." Sociology of Education, vol. 60, no. 1, 1987, pp. 2-17. JSTOR, https://doi.org/10.2307/2112615. Accessed 16 June 2026.

Rothblatt, Sheldon. Tradition and Change in English Liberal Education : An Essay in History and Culture. Faber and Faber, 1976.

UNESCO. "Education Expenditure Indicators: What Are the Challenges Going Forward?" 7 Feb. 2024, www.uis.unesco.org/sites/default/files/medias/fichiers/2025/08/Expenditure-position-paper-2023.touse__0.pdf. Accessed 9 Apr. 2026.

Impact of Pitch Clock Implementation on Major League Baseball Engagement and Player Performance By Sajan Komanduri and Abdulla Kerimov PhD

1. Abstract

Rising game times and fan disengagement led Major League Baseball (MLB) to adopt a pitch clock in 2023. This study measures its effect on game length, fan engagement, and on-field performance across twelve pre-implementation seasons (2010–2022) and three post-implementation seasons (2023–2025), excluding 2020 for its shortened COVID schedule. From a 13-variable correlation matrix, five metrics were chosen for in-depth analysis: attendance, gross revenue, batting average (AVG), stolen bases per game, and earned run average (ERA). Each was tested with a univariate framework (interrupted time series analysis, structural break testing, and counterfactual modeling) and a multivariate framework (segmented ITSA with Newey-West standard errors, PERMANOVA, and principal component analysis). Mean game length fell from 3.04 ± 0.08 to 2.67 ± 0.03 hours. Attendance reversed a multi-year decline, rising by an estimated 1,421 fans per game. Stolen bases climbed from a mean of about 0.56 to 0.72 per game. Gross revenue stayed on its existing upward path, and neither AVG nor ERA showed a break attributable to the rule. The pitch clock improved pace of play and fan engagement while leaving the underlying offensive and defensive balance largely unchanged.

Keywords: pitch clock, Major League Baseball, pace of play, attendance, gross revenue, batting average, earned run average, stolen bases, interrupted time series analysis, structural break, counterfactual modeling.

2. Introduction

Pace of play has become a central concern for professional sports leagues competing for audience attention. In MLB, worry about game length, especially the time between pitches, produced the 2023 pitch clock (Wiles). Because baseball has long resisted structural rule changes, even a modest adjustment can affect game flow, player strategy, pitching behavior, and offensive production, which makes the pitch clock one of the most consequential rule changes the sport has seen.

Attendance, viewership, and revenue reflect how enjoyable and accessible a sport is for its audience (Kriter and Sommers), and MLB expected a shorter game to raise both attendance and revenue. This study assesses how the pitch clock affected game length, fan engagement, and player performance, and it offers a framework that officials, managers, and analysts can reuse when evaluating future rule changes.

2.1 Research Gap

Few studies have evaluated the pitch clock, and none has covered a full multi-year window. Kriter and Sommers, Wiles, and Puga et al. all rely on the inaugural 2023 season alone, which leaves open whether the changes they observed are a lasting shift or single year anomaly.

Langton reaches through 2024 and Ciuffreda pools seasons before and after the pitch clock was implemented, but neither includes 2025.

Existing work also tends to study variables one at a time rather than as a system: stolen bases and pitcher tempo (Wiles); game time, attendance, and team pitching (Kriter and Sommers); pitcher injuries (Puga et al.); batting and pitching outcomes (Langton); and hitter performance from an economic angle (Ciuffreda). None of these ties the domains together or checks whether a finding in one is supported or contradicted by the others.

2.2 Study Aim

This analysis aims to assess the pitch clock's combined effect on pace of play, fan engagement, and player performance to give a comprehensive, long-term read on whether the rule marks a real turning point for MLB.

3. Methods

3.1 Dataset

MLB season statistics from 2010 to 2025 were compiled from Baseball Reference, Baseball Almanac, and published revenue reporting (Brown); 2020 was excluded because of its COVID-shortened schedule. Thirteen variables were grouped into three domains (pace of play, on-field performance, and fan engagement/revenue) and reported as means ± standard deviation for the pre- (2010–2022) and post-implementation (2023–2025) periods.

3.2 Variable Selection and Correlation Matrix

For every pair of variables, the covariance was computed and standardized by both variables' standard deviations, giving Pearson correlation coefficients (r) on a scale from −1.00 to +1.00. Five metrics with strong but non-redundant correlations were carried forward for detailed analysis: attendance, gross revenue, AVG, stolen bases per game, and ERA.

3.3 Univariate Analyses

Each metric was examined with three complementary methods. The first, interrupted time series analysis (ITSA), is a regression approach that tests whether an intervention changes the level or the trend of a variable over time. It is modeled as $Y_t = \beta_0 + \beta_1(\text{Time}) + \beta_2(\text{Intervention}) + \beta_3(\text{Time After Intervention}) + \varepsilon_t$, where β1 is the pre-2023 trend, β2 the immediate level change when the rule took effect, and β3 the change in slope afterward; R^2 gives the share of variance the model explains.

Because a visible shift in level or slope is not by itself proof of a discontinuity, structural break testing came next. The Chow test compares a restricted model, which assumes one continuous 2010–2025 trend, against an unrestricted model that fits separate pre- and post-2023 regressions. An F-statistic built from each model's residual sum of squares indicates whether the slope, the intercept, or both changed, so a significant result points to a genuine structural break rather than ordinary year-to-year variation. Counterfactual modeling then asked what each metric

would have done without the rule: a linear regression fitted to 2010–2022 was extended through 2023–2025 to build a synthetic series, and the treatment effect is the gap between the observed and synthetic values, where a value near zero means the prior trajectory simply continued.

3.4 Multivariate Analyses

Because many baseball metrics move together, a segmented ITSA with Newey-West heteroskedasticity- and autocorrelation-consistent (HAC) standard errors estimated both the level change and the slope change at 2023 while holding pre-existing trends apart from the intervention. An exact permutational multivariate analysis of variance (PERMANOVA) tested whether the multivariate weighted average of the five metrics differed between eras, and principal component analysis (PCA) showed which metrics drove that separation. Significance was set at $p \leq 0.05$, with a Benjamini–Hochberg correction applied across the five metrics.

4. Results & Analysis

4.1 Baseline Comparison

Table 1 presents mean ± standard deviation for all thirteen variables, reported separately for the pre- and post-pitch clock periods across the three domains.

Variable	Pre-Pitch Clock (2010–2022)	Post-Pitch Clock (2023–2025)
Pace of Play		
Game length (hours)	3.04 ± 0.08	2.67 ± 0.03
Player Performance		
Offense		
Stolen bases per game	0.56 ± 0.07	0.72 ± 0.02
Runs per game per team	4.38 ± 0.22	4.49 ± 0.12
Batting average (AVG)	0.253 ± 0.004	0.245 ± 0.003
On-base percentage (OBP)	0.319 ± 0.004	0.316 ± 0.004
Defense		
Earned run average (ERA)	4.07 ± 0.22	4.18 ± 0.13
Strikeouts per game	7.87 ± 0.57	8.48 ± 0.13
Walks per game	3.10 ± 0.14	3.16 ± 0.09
Balks per game	0.03 ± 0.00	0.04 ± 0.00
Innings pitched per starting pitcher	5.68 ± 0.32	5.17 ± 0.06
Fan Engagement and Revenue		
World Series viewership	15,384,173 ± 3,070,704	13,596,667 ± 3,877,871
Regular-season attendance	29,610 ± 1,278	29,291 ± 153
Gross revenue (USD)	$9,583,333,333 ± $1,193,733,639	$12,150,000,000 ± $576,628,130

Table 1. Means ± standard deviation for all 13 variables, pre and post pitch clock.

4.2 Correlation Structure

The correlation matrix (Figure 1) documents how MLB has structurally evolved. Each season (Year) correlates almost perfectly with gross revenue (0.96) and strongly with strikeouts (0.81), while correlating negatively with innings pitched per starting pitcher (−0.92) and AVG (−0.84). Together these describe a modern game of rising commercial revenue, fewer hits, and shrinking starter workloads. Game speed relates strongly and negatively to stolen bases per game (−0.93) and balks (−0.96). Attendance correlates with innings pitched per starter (0.74) and AVG (0.67) but falls as strikeouts rise (−0.74). ERA is tightly linked to walks (0.92). The final five selected metrics each hold strong ties to distinct dimensions of the dataset while avoiding redundancy, combining indicators of fan engagement and player performance.

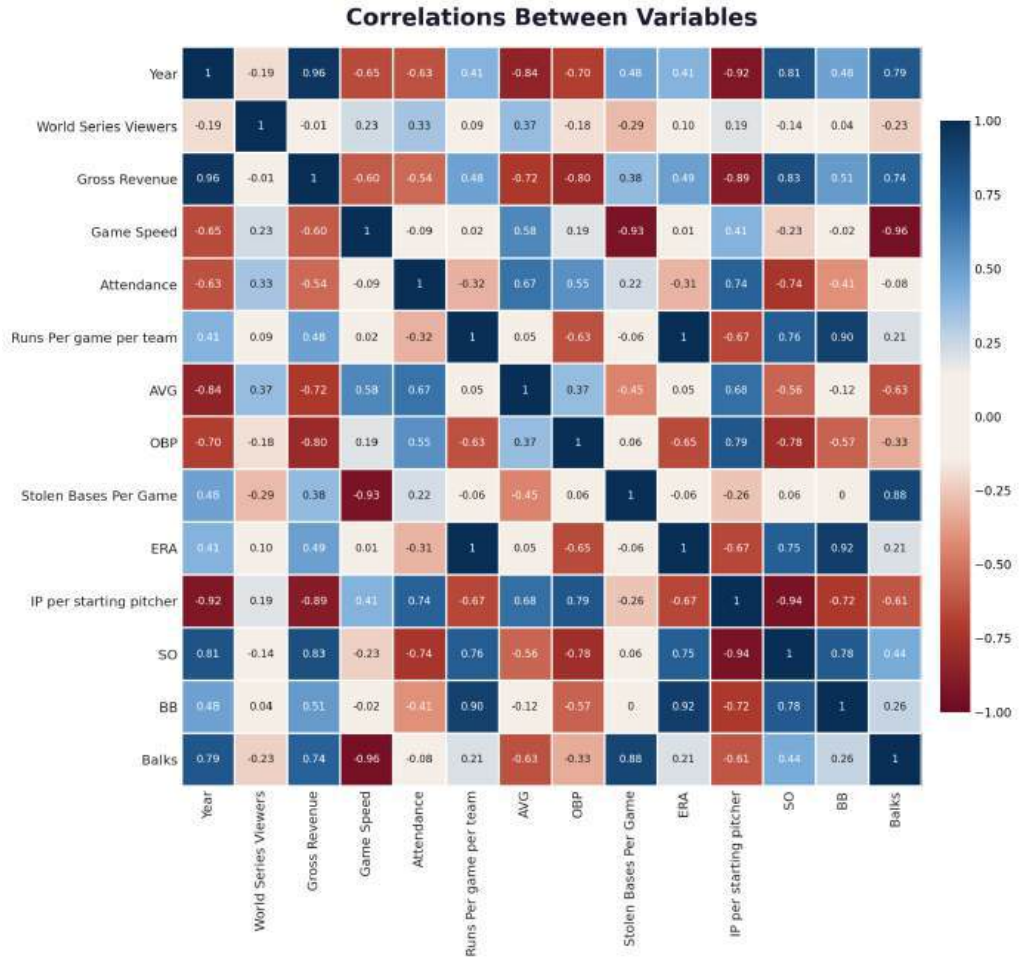


Figure 1. Correlation matrix of all 13 variables.

4.3 Pace of Play

Mean game length fell from 3.04 ± 0.08 hours in 2010–2022 to 2.67 ± 0.03 hours in 2023–2025. This is a marked reduction in duration, and game lengths cluster far more tightly around the new mean.

4.4 Player Performance

4.4.1 Stolen Bases

Stolen bases show the clearest and most durable response (Figure 2). The mean rose from approximately 0.56 ± 0.07 to 0.72 ± 0.02 per game, with the entire post-era distribution elevated. The ITSA model returned $R^2 = 0.895$, meaning the intervention model explains nearly 90% of the variability. After a steady pre-2023 decline that bottomed near 0.47 in 2019, rates jumped to 0.72 in 2023 and 0.74 in 2024. Structural break testing identified a large break at 2023 (Chow $F = 30.41$, $p < 0.001$), separating a declining pre-clock trend from a new, aggressive baseline. The counterfactual model projected continued stagnation toward 0.45 per game; the observed rate exceeded that projection by roughly 50%. The segmented ITSA confirmed an immediate increase due to the pitch clock of 0.28 steals per game (95% CI 0.22–0.34, $p < 0.001$) with a very large effect size (Hedges $g = 2.63$).

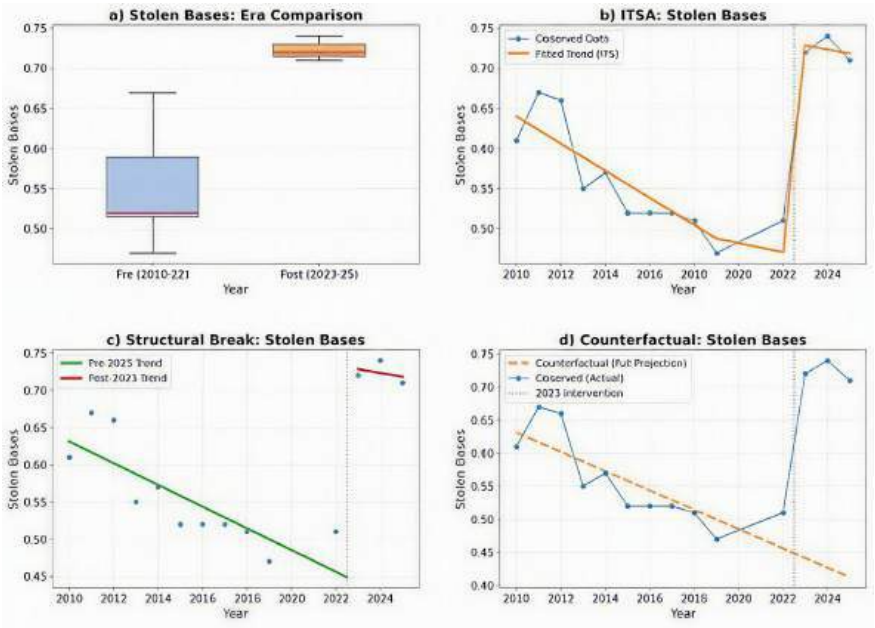


Figure 2. Univariate analysis of stolen bases per game: counterfactual model (top left), structural break analysis (top right), box plot (bottom left), and ITSA plot (bottom right; $R^2 = 0.895$).

4.4.2 Batting Average

AVG declined modestly, from a pre-pitch clock median near 0.253 ± 0.004 to about 0.245 ± 0.003, but the change is not attributable to the pitch clock (Figure 3). The ITSA ($R^2 = 0.707$) captures a consistent league wide decline since 2010 that continued unchanged through 2023–2025, and the counterfactual line projected a fall toward 0.244 that the observed values tracked closely. The structural break model produced two trend lines with similar negative slopes and no meaningful discontinuity (Chow $F = 0.07$, $p = 0.934$). Although the simple pre/post comparison is significant ($p = 0.011$), the segmented ITSA found both level and slope changes non-significant, indicating continuation of a pre-existing trend rather than an effect of the rule.

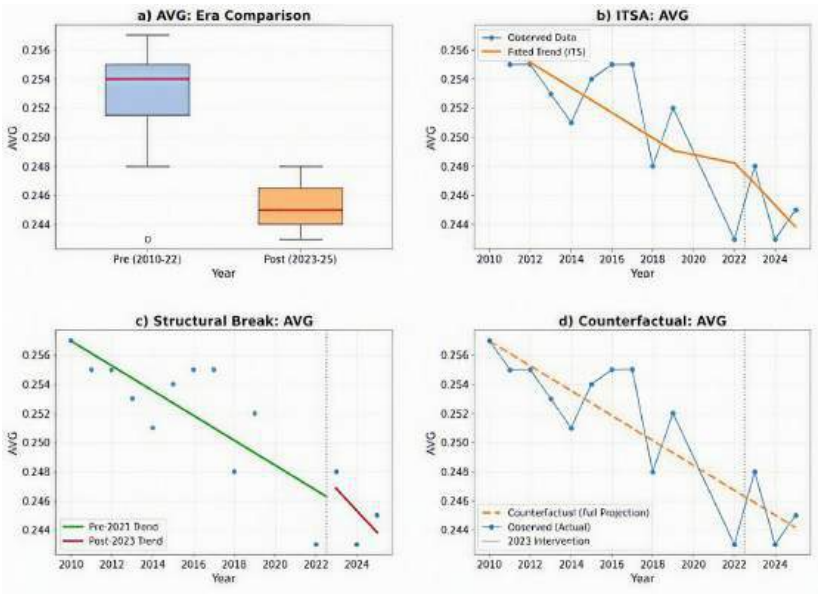


Figure 3. Univariate analysis of batting average: counterfactual model (top left), structural break analysis (top right), box plot (bottom left), and ITSA plot (bottom right; $R^2 = 0.707$).

4.4.3 Earned Run Average

ERA stabilized rather than shifted (Figure 4). The pre-era distribution was volatile, ranging from 4.07 ± 0.22 to 4.49 ± 0.12. The low ITSA R^2 of 0.344 indicates that ERA is driven largely by factors outside rule changes, including pitching technology, bullpen specialization, and rising velocity. The counterfactual model projected ERA climbing toward 4.50, whereas observed 2024 and 2025 values fell slightly below that projection. The structural break model shows a sharply rising pre-2023 trend giving way to a slightly declining post-2023 trend, but neither the Chow test ($F = 0.42$, $p = 0.667$) nor the segmented ITSA level change ($p = 0.504$) reached significance.

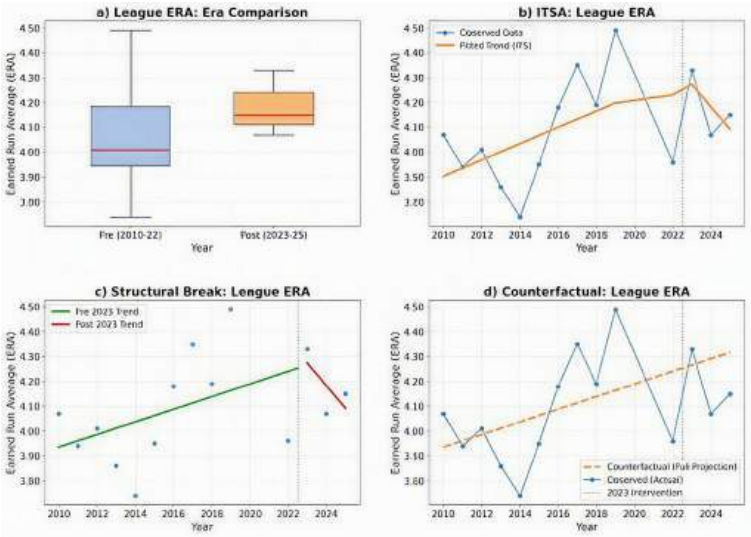


Figure 4. Univariate analysis of ERA: counterfactual model (top left), structural break analysis (top right), box plot (bottom left), and ITSA plot (bottom right; $R^2 = 0.344$).

4.5 Fan Engagement and Revenue

4.5.1 Attendance

Attendance shows a robust break at 2023 (Figure 5). The box plot alone is misleading, since the wide pre-pitch clock distribution includes early seasons with substantially higher attendance than the years immediately preceding the rule change. The ITSA ($R^2 = 0.617$) demonstrates attendance fell by roughly 800 fans per year from 2016 to 2022, reaching a historic low of 26,550, then shifted abruptly upward in 2023 to above 29,000 for the first time since 2017. Structural break testing confirmed a significant break (Chow $F = 5.68$, $p = 0.023$), and the counterfactual model projected a continued decline to about 27,000 by 2024 against a sharply rising observed series. The segmented ITSA estimated a level increase of 1,421 attendees per game ($p = 0.015$). Notably, the post-clock average of 29,291 does not exceed the long-run pre-clock average of 29,610 (simple pre/post comparison $p = 0.44$), so the effect is best interpreted as reversing a downward trend rather than exceeding historical levels.

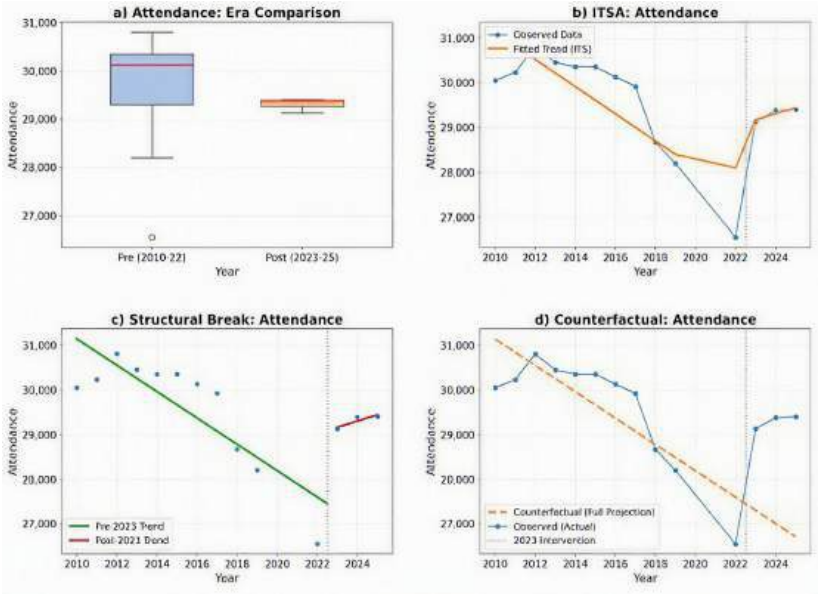


Figure 5. Univariate analysis of attendance per game: counterfactual model (top left), structural break analysis (top right), box plot (bottom left), and ITSA plot (bottom right; $R^2 = 0.617$).

4.5.2 Gross revenue

Revenue rose substantially, from a mean \$9.6B ± 1.2B pre-clock to \$12.2B ± 0.6B post-clock, but the increase reflects a pre-existing trajectory (Figure 6). Although the ITSA R^2 is high (0.928) and the structural break model shows a steeper post-2023 slope, the counterfactual model projected approximately \$12.75B by 2025 absent any rule change, essentially matching the observed value. The Chow test was non-significant ($F = 0.28$, $p = 0.760$), and while the raw pre/post difference is large ($p = 0.001$), the segmented ITSA level change was not significant once the prior trend was accounted for ($p = 0.183$). Expanded playoffs, sponsorships, and media rights plausibly account for much of the growth.

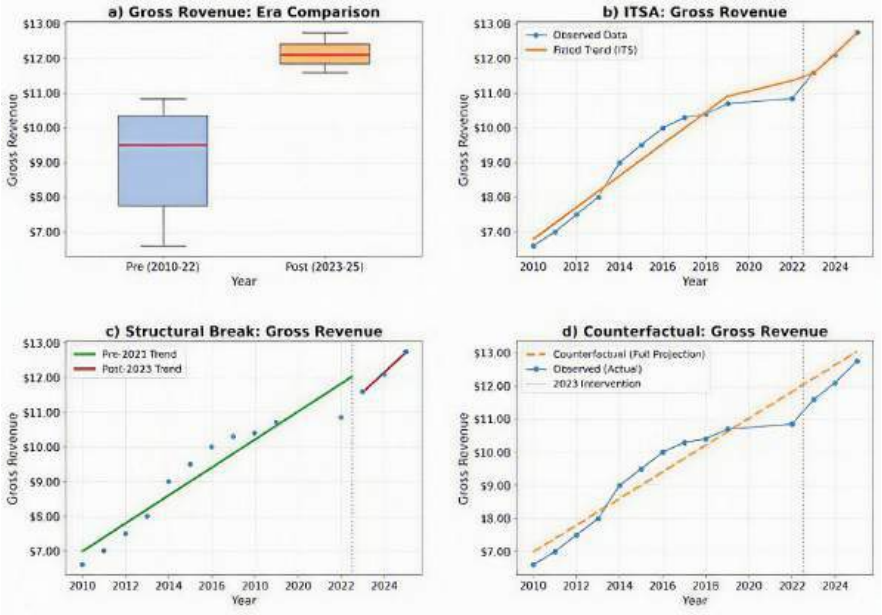


Figure 6. Univariate analysis of gross revenue: counterfactual model (top left), structural break analysis (top right), box plot (bottom left), and ITSA plot (bottom right; $R^2 = 0.928$).

4.6 Summary of Findings

Table 2 consolidates the univariate and multivariate results across all five metrics.

Metric	ITSA level change (p)	Chow F (p)	Hedges g (Welch p, BH)	PC1 loading (% PC1)	Verdict
Attendance / game	+1,421 (0.015)*	5.68 (0.023)*	−0.26 (0.437, BH 0.437)	−0.48 (23%)	Robust
Gross revenue	−$0.59B (0.183)	0.28 (0.760)	+2.16 (0.001*, BH 0.003)	+0.58 (34%)	Trend-confounded
Batting average	+0.0016 (0.573)	0.07 (0.934)	−1.78 (0.011*, BH 0.019)	−0.56 (31%)	Trend-confounded
Stolen bases / game	+0.277 (0.000)*	30.41 (0.000)*	+2.63 (0.000*, BH 0.000)	+0.23 (5%; loads PC2)	Robust
ERA	+0.123 (0.504)	0.42 (0.667)	+0.52 (0.300, BH 0.375)	+0.26 (7%)	Not significant

* $p < 0.05$. ITSA level change = immediate jump at 2023 (segmented regression, HAC SE). Chow = structural-break test in 2023. Hedges g = pre-vs-post standardized effect size; Welch p with Benjamini–Hochberg (BH) correction across the five metrics, representing each metric's univariate contribution to the significant PERMANOVA centroid shift. PCA PC1 loading = the metric's weight on the main axis separating the eras; % of PC1 (squared loading) is its share of that axis. Stolen bases contribute little to PC1 because they load on PC2 (−0.71), meaning the stolen-base surge is a separate dimension of change. Verdict: Robust = significant ITSA level change and Chow break; Trend-confounded = significant cross-sectionally (BH < 0.05) but no clean 2023 break; Not significant = neither.

Table 2. Summary of univariate and multivariate results by metric.

5. Discussion

Across a twelve-year baseline and a full three-year post implementation window, the pitch clock produced durable shifts in pace of play, stolen bases, and attendance, while leaving

batting average, ERA, and the underlying revenue trajectory largely unchanged. Game length dropped by more than twenty minutes, stolen bases settled at a new baseline near 0.72 per game, and attendance reversed a multi-year slide and stabilized. Batting average continued to fall as it had since 2010, and ERA and gross revenue stayed on their existing paths.

The metrics that moved are the ones tied most directly to the rule. Stolen bases jumped because the clock also limits pitcher disengagement, so runners take bigger leads and steal more successfully. In addition, attendance responds to a game that is shorter and has less dead time. In contrast, AVG and ERA are driven by multiple factors including, pitch velocity, bullpen specialization, defensive positioning, and hitter approach (Castrovince), and those forces carry more than a decade of momentum that a single rule was unlikely to overcome.

These findings align with and extend prior work. Kriter and Sommers found an attendance increase in the inaugural season; here that increase holds across 2023, 2024, and 2025 rather than fading as a novelty. The elevated stolen base rates reported by Wiles likewise persisted in 2024 and 2025. The revenue findings are consistent with Ciuffreda, though the counterfactual analysis suggests most of that gain would have occurred regardless.

In practical terms, the clock achieved its main goal of faster games that draw more fans, with the added effect of restoring speed on the basepaths. Where its impact was smaller, the underlying dynamics of pitching and hitting proved more resilient than a single rule could address. Decision makers weighing future interventions should distinguish between metrics that respond directly to incentive changes and those reflecting deeper structural forces evaluated over multiple seasons.

5.1 Limitations

The analysis rests on a small sample of league-average season statistics, which may mask substantial variation across teams, ballparks, and game situations. The post pitch clock window, while longer than that of prior studies, remains only three seasons, so some trends may not yet have reached equilibrium as players and coaches continue to adapt. Finally, the design identifies association rather than causation. The ball's physical composition, defensive positioning rules, and base size (Blum) all changed during the study period, so the pitch clock cannot be fully isolated from concurrent interventions.

6. Conclusion

The MLB pitch clock has meaningfully improved pace of play and fan engagement while leaving player performance largely intact. Future work should extend the time series to test whether these shifts persist or erode with further adaptation, depend on the team, ballpark, or player level to identify effects masked by league averages, and apply non-linear modeling to capture behavioral dynamics that linear approaches may miss. The lasting significance of the pitch clock will be determined not only by its immediate effects, but by whether it represents a meaningful evolution in how baseball balances tradition, pace, and the modern fan experience.

Works Cited

Blum, Ronald. “Future MLB Rules: Pitch Clocks, Shift Limits, Larger Bases.” NBC Philadelphia, 2022, www.nbcphiladelphia.com/news/sports/future-mlb-rules-pitch-clocks-shift-limits-larger-bases/3302731/.

Brown, Maury. “MLB Revenues Hit Record $12.1 Billion in 2024.” Forbes, 27 Jan. 2025, www.forbes.com/sites/maurybrown/2025/01/27/mlb-revenues-hit-record-121-billion-in-2024/.

Castrovince, Anthony. “Pitch Velocity Keeps Rising with MLB Pitchers.” MLB.com, 14 May 2026, www.mlb.com/news/pitch-velocity-keeps-rising-with-mlb-pitchers.

Ciuffreda, Katie. Ticking Time Bomb: The Economic Ramifications of MLB’s Pitch Clock. 2026. Ursinus College, Honors paper. Digital Commons @ Ursinus College, digitalcommons.ursinus.edu/bus_econ_hon/66.

Kriter, Andrew C., and Paul M. Sommers. “How Did the Pitch Clock Impact the 2023 Major League Baseball Season?” Journal of Student Research, vol. 13, no. 1, 2024.

Langton, David, Jr. An Empirical Analysis of the Pitch Clock on Major League Baseball. 2025. Clemson University, master’s thesis. TigerPrints, open.clemson.edu/all_theses/4584/.

“Major League Baseball Batting Statistics.” Baseball Reference, Sports Reference LLC, www.baseball-reference.com/leagues/majors/bat.shtml. Accessed 6 June 2026.

“Major League Baseball Miscellaneous Statistics.” Baseball Reference, Sports Reference LLC, www.baseball-reference.com/leagues/majors/misc.shtml. Accessed 6 June 2026.

“Major League Baseball Pitching Statistics.” Baseball Reference, Sports Reference LLC, www.baseball-reference.com/leagues/majors/pitch.shtml. Accessed 6 June 2026.

Puga, Troy, et al. “The Major League Baseball Pitch Clock: First Year Analysis of Pitcher Injuries.” Journal of Orthopaedics, vol. 57, 2024, pp. 133–36. https://doi.org/10.1016/j.jor.2024.06.022.

Wiles, Daniel. “Breaking Down the Pitch Clock: An Analysis of Baseball’s Big Rule Change.” Michigan Sabermetrics, 1 Mar. 2024, msabr.com/2024/03/01/breaking-down-the-pitch-clock-an-analysis-of-baseballs-big-rule-change/.

“World Series Television Ratings.” Baseball Almanac, www.baseball-almanac.com/ws/wstv.shtml. Accessed 6 June 2026.

Potential of Piezoelectric Elements on Renewable Electrical Energy Generation
By Benjamin Jetty and Henry Love

Abstract

Piezoelectric elements convert mechanical energy into electrical energy when subjected to stress. This method of electrical generation has the potential to make a substantial impact on the development of renewable energy. This research paper examines the feasibility of piezoelectric elements in roads, including necessary power harvesting circuit topologies, and quantifies the amount of energy harvested via an experimental setup. Specifically, this experiment examines the capability of piezoelectric energy harvesters to collect energy from moving vehicles. To emulate a vehicle, a skateboard was used with varying added weights ranging from 0 to 8 lbs. A test ramp was constructed to reliably start the skateboard from heights ranging from 5 inches to 25 inches. This ramp also allowed testing to determine the optimal on-track spacing between piezoelectric elements. The proposed harvesting circuit generated a peak voltage of approximately 1.5 V, commensurate with the voltage of a standard AA or AAA battery. While the total energy harvested by the system remained at or below 0.0008% of the total theoretical mechanical energy of the system, the idea shows promise for larger, more efficient piezoelectric elements that may be able to generate sufficient energy to power streetlights along roads or contribute to the nearby electrical grids while harvesting small amounts of kinetic energy from vehicles.

Introduction

Recent years have seen a significant surge in renewable energy developments, with production increasing nearly 42% from 2010 to 2020 in the United States (Center for Climate and Energy Solutions). Renewable energy is energy that originates from continuously replenished natural processes (Ciolkosz). This growth has been driven by innovations and discoveries in engineering and technology, which have improved the efficiency of traditional forms of renewable energy while encouraging an intensified search for alternative sources, such as piezoelectric energy.

Piezoelectric energy harvesting converts mechanical energy into electrical energy and has gained momentum in recent years due to its diverse applications in electrical generation through sources like roads, sidewalks, car suspension, and even body movement (APC). The expansion of research on piezoelectric materials has further accelerated interest in this field. For example, Sarker et al. conducted a bibliometric analysis of piezoelectric technologies and discovered a considerable increase in piezoelectric research being published from 2010 to 2021 (Sarker et al.).

The piezoelectric effect was first discovered by Jacques and Pierre Curie in the 1880s, and ongoing research has greatly expanded the knowledge about the workings and applications of piezoelectric materials (PI). Piezoelectric materials' crystalline structures contain crystal lattices that lack a center of symmetry (Meng et al.). When these elements are subjected to mechanical vibrations or applied stress, a net electric charge is produced due to the dipole

moment of the unit cell (Aabid et al.). Along with their capability to generate an electric potential, piezoelectric devices have numerous other advantages. They can be manufactured in multiple forms–such as patches, thin films, cylinders, and fibers, increasing their adaptability in a wide range of environments and applications (Chen et al.).

One promising application is the development of piezoelectric energy harvesters embedded in the roads to capture energy from passing vehicles. These harvesters transform mechanical energy into electrical energy that can be used to power streetlights, traffic signs, or fed into the power grid. Modern LED streetlights already consume significantly less energy compared to traditional high-pressure sodium streetlights (Davidovic & Kostic). However, they still account for nearly 20-40% of a municipality's electricity bill (Climate Group). Integrating piezoelectric energy harvesters into roadways to power streetlights could work to reduce energy costs for cities. This concept has already begun to be explored, as piezoelectric energy harvesters have been developed and experimental models have been implemented in small sections of roads (Sun et al.).

The goal of this project is to further investigate the potential of piezoelectric elements incorporated into roads to generate additional electricity. This research project includes investigating different types of energy harvesting circuits and quantifying the amount of energy harvested. The amount of total energy will be calculated using the equation $E = 1/2CV^2$, where E represents the total energy, C is the total capacitance of the circuit, and V is the maximum voltage, as capacitors are used to temporarily store the produced voltage. In addition, trade-offs among circuit topologies, piezoelectric element placements, and vehicle ideal speed and weight will be discussed.

Circuit Topologies

In this project, LTSPICE, a circuit simulation software, was used to experiment with circuit types to determine which layout produces the most energy when paired with a piezoelectric element (Analog Devices). An ideal sinusoidal voltage source was used to emulate a piezoelectric element's output voltage. While this model has limitations and continuously provides input to the rectifier, it gives a time-varying signal that can be used to compare the two system architectures. Also, due to the time-varying nature of traffic, the piezoelectric element creates a time-varying voltage/current waveform as vehicles drive over it. In some cases, the piezoelectric elements generate zero voltage; however, when a wheel passes over the element, a voltage/current spike is generated. A capacitor stores these abrupt current spikes and slowly discharges them into a load. In addition, because of the time-varying waveform, to generate a consistent energy output, a rectifier is used to convert this spike to a DC/constant waveform. A rectifier is an electrical device used to convert alternating current (AC) into direct current (DC) by allowing a current to flow through the device in one direction only (Dynapower).

Common rectifier circuits include half-wave and full-wave rectifiers, which both utilize a capacitor(s). The half-wave rectifier, as shown in Figure 1 (left), converts only half of the total output of the piezoelectric harvester: one side of the sinusoidal wave. However, because only one

diode is used, the forward voltage drop (reduction of current through a diode) is lower, and less voltage is consumed during conversion. The remaining output is quickly stored in the capacitor and then discharged into the load resistor. On the other hand, in a full-wave rectifier, as shown in Figure 1 (right), both sides of the sinusoidal wave are converted to DC, but due to there being four diodes, the voltage drop across the rectifier is greater, which limits the available output voltage. However, since the full-wave rectifier operates on both polarities of the sinusoid, there is less ripple on the output voltage compared to the half-wave rectifier. This voltage is then quickly stored in the capacitor and then discharged into the load resistor. Ultimately, a full-wave rectifier was selected because it generated a voltage with a lower ripple.

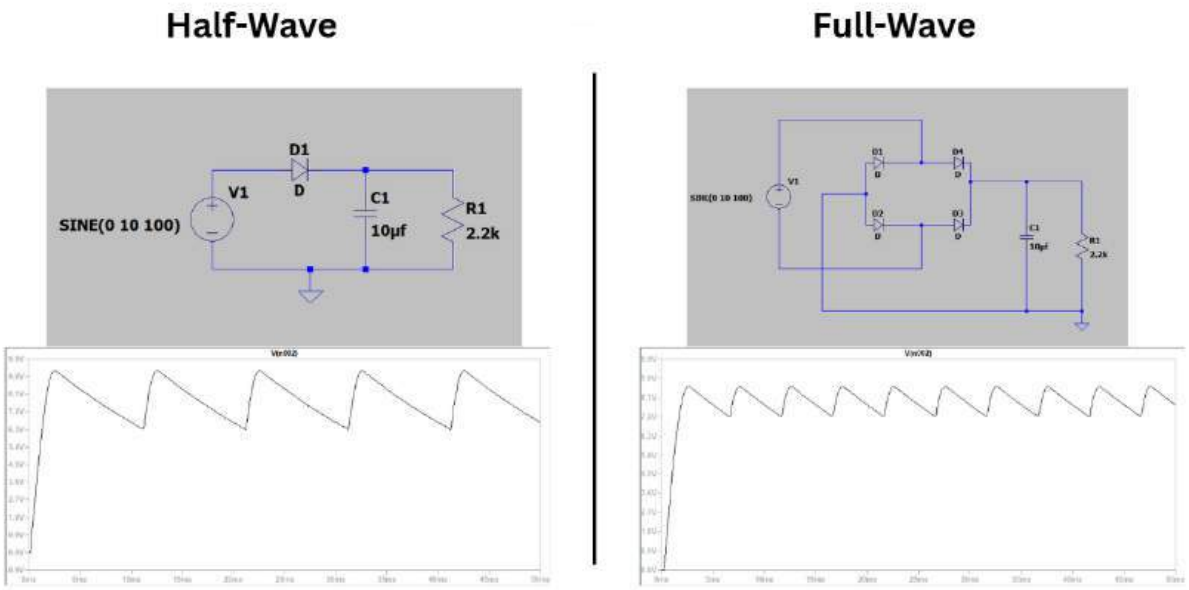


Fig. 1. Half-wave rectifier circuit (top left) with output voltage (voltage across R1) plotted in (bottom left), one side of the sinusoidal input source is rectified through D1, then connected in parallel to a 10 µF capacitor (C1) and a 2.2 kΩ load resistor (R1). Full-wave rectifier circuit (top right) with output voltage (voltage across R1) plotted in (bottom right). Both sides of the sinusoidal input source are rectified through D1-D4, then connected in parallel to a 10 µF capacitor (C1) and a 2.2 kΩ load resistor (R1).

Another simulated configuration compared the performance of a single diode to that of multiple diodes, where both cases are shown in Figure 2 below. Figure 2 (left) shows a single diode circuit that is shared between multiple piezoelectric elements, whereas Figure 2 (right) has a separate diode per individual piezoelectric element. In this test, a half-wave rectifier was simulated to ease the setup. In the simulation, the circuit with multiple diodes had a voltage with less ripple pass through the resistive load than the single diode circuit. Moreover, when experimenting with different time delays to replicate a vehicle traveling over individual piezoelectric generators, the multiple diode circuit greatly outperformed the single diode circuit, which failed to work without low-value resistors connecting the energy sources. This was caused

by the independent ideal sinusoidal sources, which cancel each other out when they are connected in parallel.

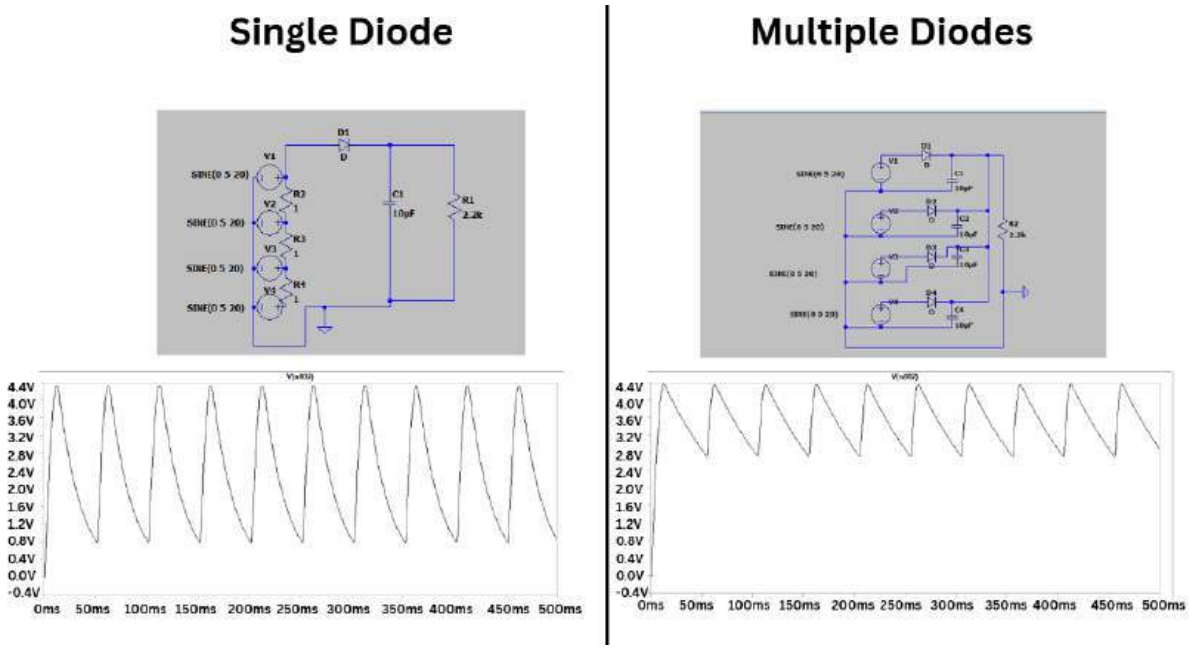


Fig. 2. Single diode (top left) with voltage across R1 plotted (bottom left), 4 sinusoidal input sources pass through one diode (D1) and are connected in parallel to a 10 µF capacitor (C1) and a 2.2 kΩ load resistor (R1). Multiple diodes (top right) with voltage across R1 plotted (bottom right), 4 sinusoidal input sources pass through their own diode (V1-D1, V2-D2, V3-D3, V4-D4) and are connected in parallel to their own 10 µF capacitor (C1-C4) and a 2.2 kΩ load resistor (R1).

Harvester Schematic

Based on the simulations above, the final circuit designed and tested was a full-wave rectifier that combined multiple piezoelectric elements, each with multiple diodes, as shown in Figure 3 below. The positive and negative wires of each piezoelectric element were connected and passed through a load resistor. Each piezoelectric element had its own diode and its own capacitor. This circuit was chosen due to the increased electrical output. Since the power present in a circuit is equal to $P = IV$, the greater current and voltage potential of the selected design resulted in more total energy harvested.

In the circuit, mechanical stress creates an alternating current produced from the crystals within the piezoelectric element (Ballas). The current travels to the full-wave rectifier, where it changes from AC to DC, which can then be stored in the capacitor. The capacitor is used to quickly store the surge of energy produced by the piezoelectric element, which is then discharged to the resistor load to simulate a practical load that would be present in an energy-harvesting application such as a streetlight.

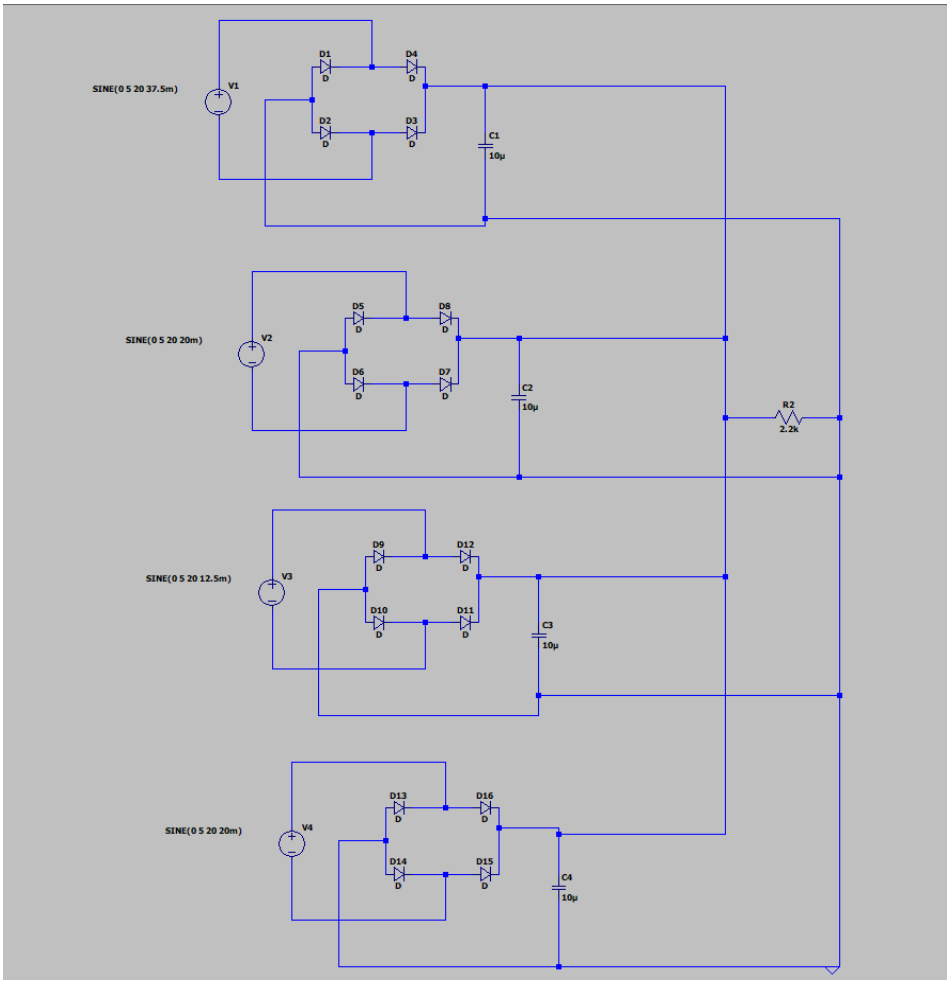


Fig. 3. Final circuit model: full-wave rectifier on each sinusoidal source, each with its own 10 uF capacitor, connected in parallel to a 2.2 kΩ resistor (R2) to emulate a load.

Preliminary Tests

Three different sizes of piezoelectric elements were used: 41 mm, 50 mm, and 65 mm diameters. Through testing, I determined that the size of the piezoelectric elements affected the voltage obtained, along with the location of the impact. Larger piezoelectric elements generally produced higher voltages during testing. In addition, more centered hits resulted in a larger pulse of voltage. Therefore, based on these metrics, the 65 mm piezoelectric element performed the best, and the 41 mm performed the worst. However, the size also greatly impacted the cost of the piezoelectric elements, with larger piezoelectric devices costing more. Despite this, the 65 mm piezoelectric disc was chosen for the final configuration.

The size of the capacitor and the resistor was tested through various simulations on LTSPICE. An increased capacitor value led to a smoother DC with less ripple, whereas a decreased capacitor value had the opposite result, with more ripple between the charges from the sinusoidal source. On the other hand, increasing the resistor value led to less current passing through the resistor as the capacitor drained more slowly, producing less ripple. Decreasing the

resistor value led to a faster-draining capacitor and thus more ripple on the output. The resistor value gave insight into how many devices the harvester could power, but with only hypothetical, perfect sinusoidal sources, the final output voltage remained to be found.

Methods

To quantify the energy and voltage generated by piezoelectric elements, an experiment was created to simulate a vehicle passing over the piezoelectric energy harvester. A 4.5 lbs skateboard was used to emulate a vehicle with the ability to add weight to the skateboard platform. To repeatedly test different vehicle velocities, an adjustable test ramp was constructed, see Figure 4 below, with guides to ensure the skateboard traveled along a consistent path to produce controlled test results. The ramp was able to move at 5-inch intervals above the surface level of the piezoelectric elements, starting at 5 inches and ending at 25 inches. The guides along the surface board were constructed to ensure consistent test trials and to clamp down on the piezoelectric devices to prevent movement during test trials. The guides could also be raised to allow the piezoelectric devices to be adjusted to test if the spacing had any effect.

Ten piezoelectric buzzers were placed on the board, each with its own full-wave rectifier and 10 μF capacitor. The output of each rectifier was then connected to the central breadboard, where it passed through a load 2.2 kΩ resistor to emulate a power draw. The voltage across the resistor was connected to an Arduino, which measured the waveform. Three variables were tested on the piezoelectric elements: the spacing between elements (5, 10, 15, and 20 cm), the additional weight on the skateboard (0, 2, 4, 6, and 8 lbs), and the height of the ramp (5, 10, 15, 20, and 25 in) to affect the speed of the skateboard. These results are compared to determine what is the best configuration of piezoelectric elements and the ideal location for their placement.

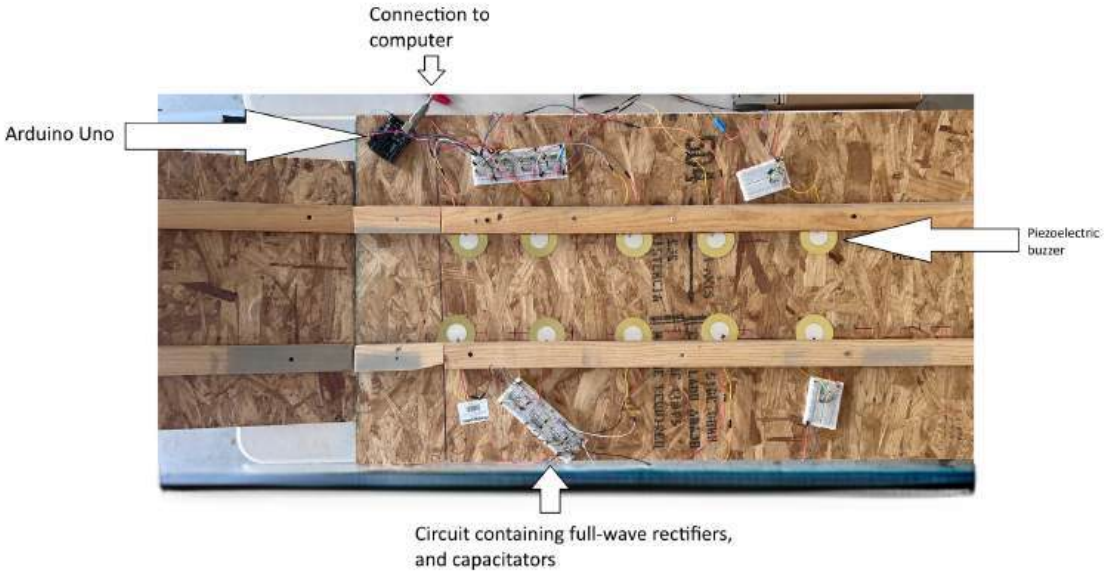


Fig. 4. Experimental setup. An overhead view is depicted on top, and a side view is depicted in the middle, with part numbers listed in Figure 5

Parts	**Part number**
BUZZER ELEMENT STD 520HZ 65MM	AB6505B
CAP ALUM 10UF 20% 50V RADIAL TH	106CKH050M
DIODE SCHOTTKY 100V 100MA DO35	BAT41

Fig. 5. Part numbers of the circuit elements used in the experiment.

Data

During the first set of tests, 3 trials were run at each spacing interval with 0lb on the skateboard, and the ramp was elevated 10 inches. The results of this test, shown in Figure 6 below, indicate that when placed 15 cm apart (gray line), the piezoelectric energy harvester produced the highest maximum voltage. As a result, this spacing was used in the set of tests. The next set of tests consisted of three trials at each height and weight category. The results are depicted below in Figure 7. This data represents the average of three trials for each weight and speed category. The voltage is depicted on the y-axis, and time in milliseconds is shown on the x-axis. This data was then used with the equation $E = 1/2CV^2$ to determine the electrical energy generated from each trial.

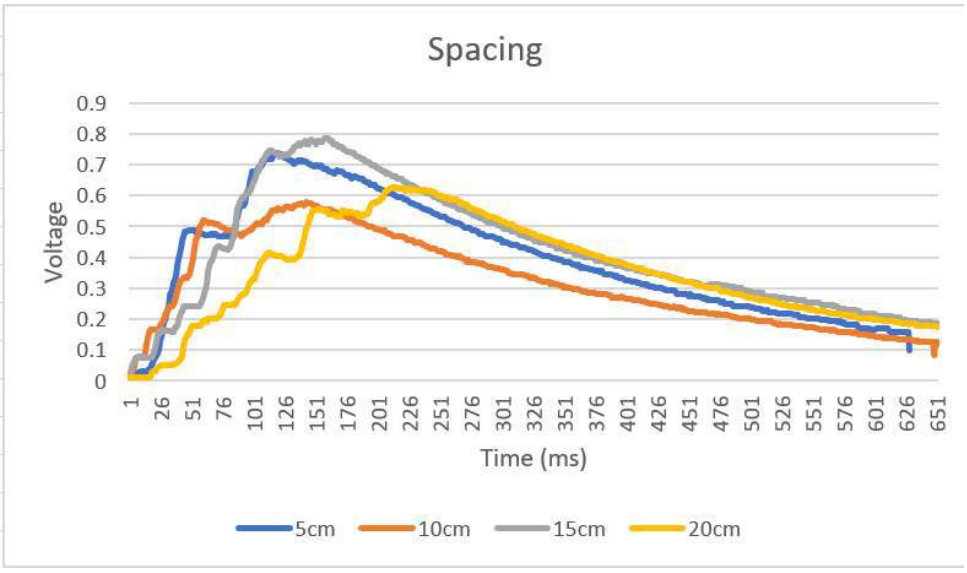


Fig. 6. Voltage generated from each piezoelectric spacing trial. Each line represents a different spacing interval between the piezoelectric elements. Blue - 5cm, orange - 10cm, gray - 15cm, yellow - 20cm.

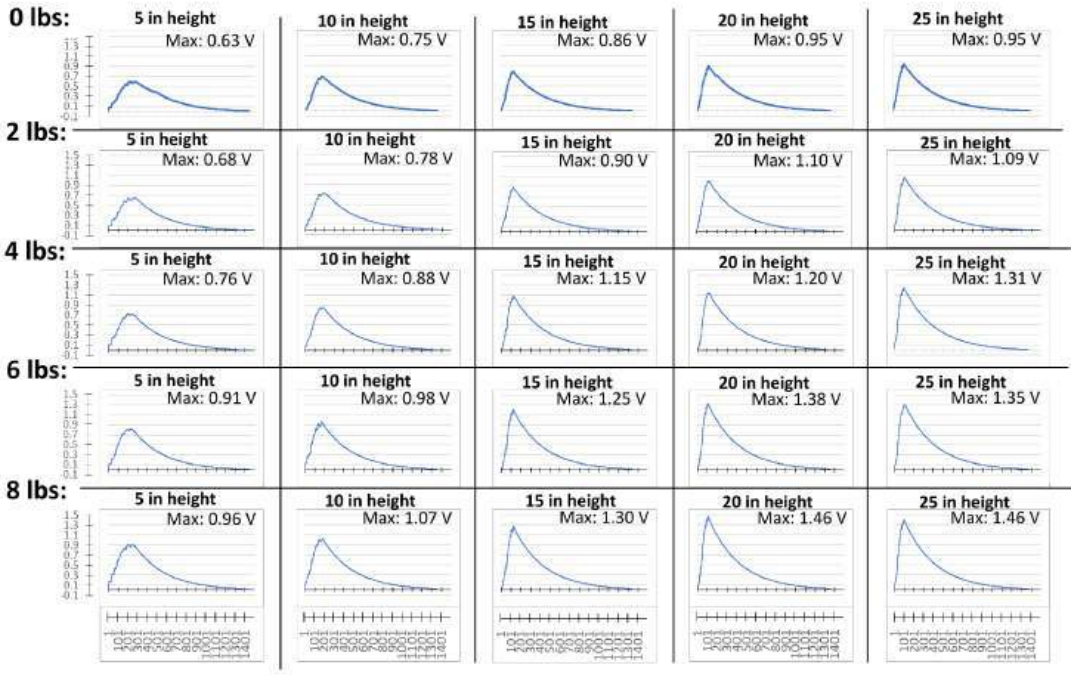


Fig. 7. The average result of each test, with voltage on the y-axis (0 to 1.5 V) and time in milliseconds plotted on the x-axis. The x-axis range is from 0 to 1500ms.

A link to all the raw data collected during experimentation can be found here: https://docs.google.com/spreadsheets/d/1Xftnmk4xpAL9oGS1VddYU4S6u3XXmXWc/edit?usp=sharing&ouid=111116709037398079734&rtpof=true&sd=true

Data Analysis

The goal was to calculate the total electrical energy obtained from each test using the equation $E = 1/2CV^2$ and comparing it to the maximum potential energy of the skateboard from the top of the ramp $E = mgh$. The first equation used a capacitor value of 10 μF for C and used the peak voltage, as annotated in each plot, for V. The results of the data, as depicted in Figure 8 below, show that a small percentage of the total energy of the skateboard was converted into electrical energy, approximately 5 x 10^{-4}%. The highest percentage obtained was 8 x 10^{-4}% from the lightest and slowest skateboard, as shown in Figure 8 below. This low amount of total energy was likely due to the piezoelectric elements used, as they were small and low-cost components. The differences in the percentages are likely due to the skateboard bouncing over the piezoelectric elements at higher speeds, as the elements were placed on the wood rather than embedded in it to prevent surface bumps. However, more voltage was generated at increased speeds and heavier weights, which could be advantageous because less kinetic energy is converted to electrical energy, allowing vehicles to travel over the harvesters with little effect on their speed and gas usage.

	5in	**10in**	**15in**	**20in**	**25in**
0lb	0.019845 mJ	0.028125 mJ	0.036 mJ	0.045125 mJ	0.045125 mJ
2lb	0.02312 mJ	0.03042 mJ	0.0405 mJ	0.0605 mJ	0.05941 mJ
4lb	0.02888 mJ	0.03872 mJ	0.06613 mJ	0.072 mJ	0.08581 mJ
6lb	0.04141 mJ	0.04802 mJ	0.07813 mJ	0.09522 mJ	0.09113 mJ
8lb	0.04608 mJ	0.05725 mJ	0.0845 mJ	0.106mJ	0.106mJ

Fig. 8. Height of ramp in the first column and weight of the skateboard

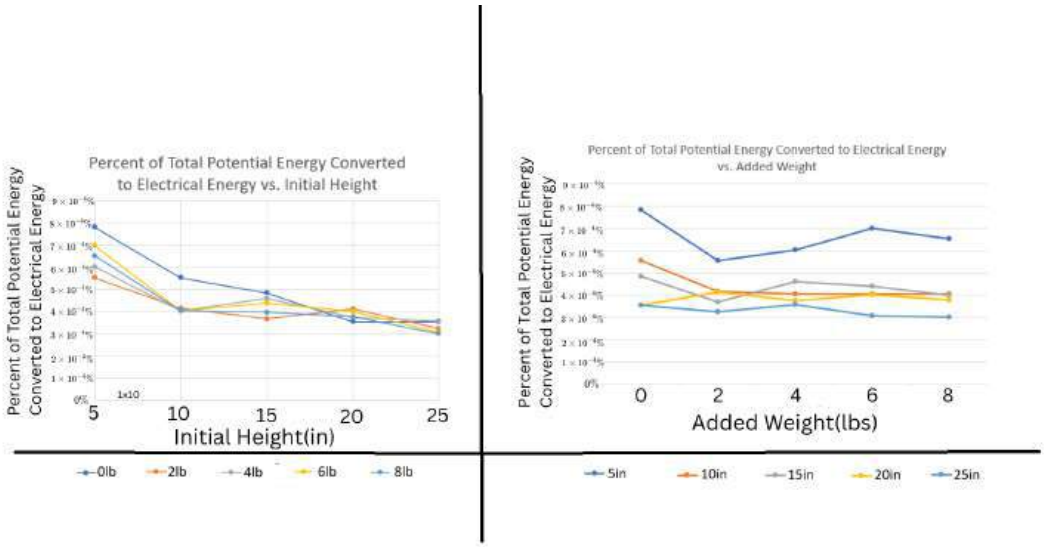


Fig. 9. On the left, Percent of Total Potential Energy Converted into Electrical Energy vs. Initial height, the colors display different weights. On the right, Percent of Total Potential Energy Converted into Electrical Energy vs. Added Weight, the colors display different heights. Both graphs display a y-axis with values in increments of 1 x 10^{-4}%.

Conclusion

These results demonstrate that although the amount of energy generated from the piezoelectric harvesters was relatively small, the concept still shows significant potential. Theoretically, under ideal conditions, if the relationship between voltage and weight remains constant, a rough estimate assuming linear scaling suggests that an average car (1300 kg) could generate around 1.5488 J at a time. If 500 vehicles run over the harvester in a day, the harvester could generate 774.4 J. These numbers assume that the vehicles are traveling at slow speeds, so at faster speeds, these numbers could be increased, based on the experimental trends observed (i.e., faster skateboard velocity produced higher peak voltage). However, this does not take into account energy loss through batteries or non-ideal conditions.

Moreover, implementing this on real roads poses challenges, and several things need to be improved in order to establish it on actual roadways. The piezoelectric energy harvesters need to be embedded in the concrete or asphalt of the roads rather than being placed on top of roads to prevent rough surfaces. In addition, the durability needs to be increased to enable them to survive the weather and the vehicles traveling over them. Piezoelectric elements can be relatively expensive, and the larger sizes needed to effectively capture the same proportion of total energy from larger vehicles dramatically increase the price of piezoelectric elements, creating a potential barrier for low-wealth areas. The diodes and capacitors also need to be changed to be able to handle the increased amounts of voltage.

Currently, this design captures a very minuscule percentage of the total kinetic energy of vehicles. To be worth the cost of implementing piezoelectric energy harvesters in the roads, the harvesters need to be more efficient. However, this project still indicates the future potential of piezoelectric energy harvesters to provide another source of renewable energy. By being placed on highways, fast-speed roads, or places where vehicles need to slow down, piezoelectric harvesters could capture large amounts of energy from traffic in order to power nearby lights without decreasing gas mileage. Piezoelectric harvesters would function best at high-density areas like cities, where there are many vehicles driving over each harvester. In contrast, low-density locations like rural areas would be an unfavorable place for piezoelectric harvesters, as the costs would very quickly outweigh the benefits received from them. The areas where the most improvement could be made on this topic involve more efficient piezoelectric elements that could capture more of the total energy. Overall, this research supports the future development of piezoelectric roads.

Works Cited

Aabid, A., et al. "A Systematic Review of Piezoelectric Materials and Energy Harvesters for Industrial Applications." Sensors, vol. 21, no. 12, 2021, p. 4145. https://doi.org/10.3390/s21124145.

Ballas, R. G. "The Piezoelectric Effect – an Indispensable Solid State Effect for Contemporary Actuator and Sensor Technologies." Journal of Physics: Conference Series, vol. 1775, no. 1, 2021, p. 012012. https://doi.org/10.1088/1742-6596/1775/1/012012.

Center for Climate and Energy Solutions. "Renewable Energy." Center for Climate and Energy Solutions, Nov. 2017, https://www.c2es.org/content/renewable-energy/.

Chen, B., et al. "Flexible Piezoelectrics: Integration of Sensing, Actuating and Energy Harvesting." npj Flexible Electronics, vol. 9, no. 1, 2025, https://doi.org/10.1038/s41528-025-00432-5.

Ciolkosz, Daniel. "What Is Renewable Energy?" Penn State Extension, 20 Apr. 2019, https://extension.psu.edu/what-is-renewable-energy.

Climate Group. "LED." The Climate Group, 2025, https://www.theclimategroup.org/node/221.

Davidovic, M., and M. Kostic. "Comparison of Energy Efficiency and Costs Related to Conventional and LED Road Lighting Installations." Energy, vol. 254, 2022, article 124299. https://doi.org/10.1016/j.energy.2022.124299.

Dynapower. "Rectifiers: Nearly Everything You Need to Know." Dynapower, 2023, https://dynapower.com/rectifiers-nearly-everything-you-need-to-know/.

"LTspice Information Center." Analog Devices, 2024, https://www.analog.com/en/resources/design-tools-and-calculators/ltspice-simulator.html.

Meng, Y., G. Chen, and M. Huang. "Piezoelectric Materials: Properties, Advancements, and Design Strategies for High-Temperature Applications." Nanomaterials, vol. 12, no. 7, 2022, article 1171. https://doi.org/10.3390/nano12071171.

Physik Instrumente (PI). "Fundamentals of Piezo Technology." Physik Instrumente (PI), n.d., https://www.pi-usa.us/en/expertise/technology/piezo-technology/fundamentals.

Sarker, M. R., et al. "A Bibliometric Analysis of Low-Cost Piezoelectric Micro-Energy Harvesting Systems from Ambient Energy Sources: Current Trends, Issues and Suggestions." Micromachines, vol. 13, no. 6, 2022, article 975. https://doi.org/10.3390/mi13060975.

Sun, J.-Q., T.-B. Xu, and A. Yazdani. "Ultra-High Power Density Roadway Piezoelectric Energy Harvesting System." California Energy Commission, 1 June 2023, https://www.energy.ca.gov/publications/2023/ultra-high-power-density-roadway-piezoelectric-energy-harvesting-system.

"Top Uses of Piezoelectricity in Everyday Applications." APC International Ltd., 13 Feb. 2015, https://www.americanpiezo.com/blog/top-uses-of-piezoelectricity-in-everyday-applications/.

Safety and Efficacy of Topical Acne Ingredients: A Narrative Review By Diya Rishi

Abstract

Acne is one of the most common skin disorders, affecting the majority of adolescents and adults at some point in their life. With the sheer variety of acne treatments available, many choose products based on the belief that natural ingredients are gentler on the skin, while synthetic ingredients are more effective at reducing acne. This narrative review compared evidence on the efficacy and tolerability of selected botanical ingredients, including tea tree oil and aloe vera, with conventional agents including benzoyl peroxide, salicylic acid, and azelaic acid. Literature available through July 2026 was identified through PubMed, Google Scholar, JSTOR and the reference lists of relevant reviews and guidelines, with priority given to controlled human studies reporting lesion counts, acne severity, or adverse events. By reviewing naturally derived and synthetic topical acne ingredients, this review offers a comprehensive guide to understanding the products that balance efficacy and safety.

Definitions/Keywords

Acne Vulgaris: a common inflammatory disease characterized by comedone and pustules driven by sebum and excess sebum production

Over-the-Counter Topicals: OTC products applied directly on the skin without the need for a prescription

Natural Ingredients: plant-based or herbal substances like tea tree oil and aloe vera

Synthetic Ingredients: chemical formulated compounds such as benzoyl peroxide, salicylic acid, and azelaic acid

Introduction

Biologically, acne is an inflammatory skin condition characterized by excess sebum production, clogged pores, and bacterial growth leading to the growth of comedones, pustules, and lesions (Sutaria et al.). Globally, acne vulgaris is estimated to affect around 9.4% of the general population, being the eighth most prevalent disease in the world. Acne is especially common during puberty where hormonal changes occur, affecting 85% to 100% of adolescent males and 83% to 85% of adolescent females between the ages to 16 and 17 . In addition to the physical impacts, individuals with this condition may experience a lack of self-esteem or confidence due to the visible nature of the condition (Jisa et al.).

Standard treatment guidelines are established from evidence-based clinical research and guidelines from dermatological authorities. Today, over-the-counter topicals including benzoyl peroxide, salicylic acid, and azelaic acid are commonly used as the first step in managing acne (Reynolds et al.). Even though these are made in labs so companies can produce them in large batches, those synthetic products are derived from nature (Ramsis et al.). However, they are still synthetic ingredients, since they are made in a lab for consistency, compared to natural ingredients which are raw complex mixtures and extracts taken directly from plants.

Despite the efficacy of these products, there are also limitations of current treatments (Sutaria et al.). Some topicals create issues like irritation, dryness, or sensitivity, which may lead to a lack of consistency using those products. In earlier years, public concerns were over cosmetic safety in usual synthetic treatment plans, which ended up sparking the "clean ingredient" movement (Wanner and Nathan). "Clean" products are usually described as products made without any chemicals like parabens or synthetic fragrances. Instead, those products focus on simple raw ingredients. Natural ingredients in acne treatment include tea tree oil, aloe vera, and other medicinal plants and raw extracts, while synthetic ingredients are the usual acne treatment. Comparing natural and synthetic ingredients is important because it can address limitations. This trend has led many people to look at natural alternatives compared to the typical dermatological treatments to avoid common effects like skin irritation.

The current research landscape for natural and synthetic ingredients differs. Synthetic ingredients are backed by a large number of resources and human clinical trials, while there is much limited and smaller amount of studies for natural ingredients. To analyze the current research, the paper will evaluate and compare the studies based on efficacy, tolerability, and the study quality.

The purpose of this review is to evaluate and compare natural and synthetic topical ingredients within the context of their efficacy and safety in treating acne, thus informing others of the current evidence-based findings, providing a critique of the field, raising new questions for future works, and summarizing a cross section of current knowledge about the topic. This narrative review bridges the gap between the popular natural alternatives against many clinical studies for synthetic topicals and proposes a potential solution for future research and treatment plans.

Methods

Search Strategies

Multiple search engines and academic databases were used to identify relevant literature on acne vulgaris and its current treatments, including *PubMed*, *Google Scholar*, and *JSTOR*.

Additional sources were gathered from peer-reviewed journals such as the *Journal of American Academy of Dermatology*, the *Journal of Cosmetic Dermatology*, *ScienceDirect*, *Frontiers*, and *Wiley*. Other information from medical clinics or organizations like the *Mayo Clinic* were also utilized.

Searches were conducted using keywords like *acne vulgaris, natural acne remedies, tea tree oil acne, aloe vera acne, synthetic acne treatments, benzoyl peroxide, salicylic acid, and azelaic acid.*

Inclusion and Exclusion Criteria

Studies were included if they focused on acne vulgaris, evaluated over-the-counter products, and examined either natural or synthetic ingredients. Sources published between

2011-2026 were included, with the exception of a couple of older sources that were used solely for background information and one study.

Studies were excluded if they focused on severe dermatological issues unrelated to acne, targeted very specific populations or uncommon skin conditions, were not directly related to acne treatment, or primarily examined on prescription treatments.

Selection Process

Articles were first screened based on their abstracts to determine their applicability, particularly in relation to the efficacy of natural and synthetic ingredients for acne treatment. Selected studies were then further reviewed and read to evaluate their relevance to the research question. Additionally, the credibility of each source and the reliability of the author were considered, including an assessment of their expertise in the field.

Results

Natural Ingredients

While natural remedies are more popular because consumers believe they are safer and less likely to cause side effects, clinical studies and evidence show mixed results when it comes to the efficacy in clearing acne (Institute). In a 12-week randomized, double-blind, placebo-controlled trial, a topical botanical treatment of basil leaf extract and berberine root extract significantly improved acne severity compared to placebo, reducing inflammatory and non-inflammatory lesions, while overall improving skin appearance (Draelos et al.). For three months participants were split into groups at random where some got the real treatment and others got a placebo. It was blinded so no one knew who had gotten which treatment, which lowered the chances of bias. However, there were no significant differences in oil or sebum production between the groups. Also, it was found that a natural cream extract made of 20% propolis, a substance produced by bees, 10% aloe vera, and 3% tea tree oil, reduced acne lesions and redness by 63.7% acne compared to the other option, synthetic 3% erythromycin cream, which decreased acne lesions by 46.5% (Torjesen et al.). Additionally, it was reported that natural products such as tea tree oil and aloe vera usually produce fewer side effects like redness or irritation than synthetic formulations (Ramsis et al.). In the clinical trials that were reviewed by Ramsis et al., side effects were limited to itching minor stinging, in the study with 5% tea tree oil compared to the aloe vera with tretinoin, which was conducted on both male and female adolescents from ages 12 to 30.

Tea tree oil is one of the most well studied natural topical ingredients. This ingredient effectively reduces acne by decreasing bacterial growth and inflammation, although its effects are overall slower than other treatments. While many studies researched reductions in acne vulgaris, the relationship between tea tree oil use and acne improvement was not the same across all trials. Some reported side effects include skin irritation, itching, dryness, and allergic reactions, and the side effects from tea tree oil application were minimal and some formulations

provide rather hydrating benefits. (Mayo Clinic Staff). In randomized controlled trials testing efficacy and safety of tea tree oil for human health, adverse effects were very mild and limited to benign skin conditions, with more severe reactions associated with 25% or greater use of tea tree oil in general (Kairey et al.). A 2023 review of 46 controlled clinical trials concluded that tea tree oil demonstrated a potential as acne treatment, substantiating its antimicrobial and anti-inflammatory effects, although the authors noted that additionally high-quality studies are still needed to strengthen the evidence base (Reynolds et al.). In one such 12-week randomized trial with 100 patients, the combination of tea tree oil and 0.1% adapalene outperformed adapalene alone in reducing lesion count. Because acne lesions had lowered throughout the entire study, treatment was a success for 71.7% patients receiving the tea tree oil and adapalene combination, while the adapalene alone decreased by 6.4%. Dryness was the most common side, affecting 68.8% in the tea tree oil group, while there was 53.2% in the adapalene-only group (Najafi-Taher et al.) (Table 1). In a different randomized, single-blind clinical trial involving 124 patients with mild-moderate acne both 5% tea tree oil gel and 5% benzoyl peroxide reduced inflammatory and non-inflammatory acne lesions. However, benzoyl peroxide produced faster improvement, while tea tree oil caused substantially fewer side effects, like skin irritation, redness, and dryness from the benzoyl peroxide (Bassett et al.) (Table 1). Additional evidence from a 12-week pilot study using a tea tree oil facial wash and gel treatment found that the average lesion counts decreased by 23.7 to 10.7 lesions at the end of the experiment, which is around a 54% reduction, where 79% of people achieve success in decreasing their acne generally (Malhi et al.) (Table 1). These studies contribute to the research goal of finding the efficacy of natural ingredients through clinical data and studies on tea tree oil.

Findings regarding the efficacy of aloe vera in acne vulgaris also differed greatly. Aloe vera is solely beneficial for its soothing properties rather than as acne treatment, according to a clinical review on botanical therapies (Bosslett et al.). However, it found that aloe vera contains numerous biologically active compounds like acemannan, which contains wound-healing, anti-inflammatory, and antimicrobial properties (Nascimento et al.). Since acne development is associated with bacterial overgrowth, excess sebum production, and inflammation, aloe vera can help reduce acne severity by controlling acne-causing bacteria (Pal et al.). Aloe vera is composed of 96% water and 4% dry matter, which explains the hydrating and skin-repairing effects. These properties have also made aloe vera a part of non-drug acne treatments to improve skin healing while minimizing irritation (Zhong et al.). Other evidence suggests that aloe vera may be more effective when combined with synthetic treatments rather than alone. In an 8-week randomized clinical trial, topical tretinoin with aloe vera gel and topical tretinoin with placebo. Both groups showed significant reductions in acne lesions, but the aloe vera group experienced greater reduction in lesion counts between weeks four and eight and achieved an overall acne improvement by 78.8%. Although side effects remained common, affecting 76.8% of participants, the aloe vera group experiences significantly less redness than the placebo group ($p=0.046$), meaning that aloe vera may help reduce some of the irritation associated with synthetic ingredients (Hajheydari et al.) (Table 1).

Synthetic Ingredients

Benzoyl peroxide is one of the most commonly used over-the-counter acne treatments, and works by killing acne-causing bacteria while also helping to unclog pores (Mayo Clinic Staff). When used correctly, it is generally considered safe and effective, with expected side effects including dryness and redness. Benzoyl peroxide is effective for inflammatory lesions and pustules. Despite the benefits of this treatment, in a double-blind, randomized, controlled trial, Kaeswanit et al. compared 2.5% benzoyl peroxide alone with a separate mixture containing 5% niacinamide in patients with mild to moderate acne. Both treatments significantly reduced acne lesions, although the combination mixture reduced noninflammatory and inflammatory acne lesions by 51%, while the benzoyl peroxide alone decreased lesions by 34% (Kaewsanit et al.). The study also reported improvement is sebum production and overall severity of the acne for both mixtures, which is further why benzoyl peroxide is such a common acne medication. However, there are also concerns about the long-term safety of this medication. In an animal study, researchers applied benzoyl peroxide to mice twice weekly for 52 weeks and found that high doses lead to the development of skin tumors over time after prior exposure. Further research found overuse and exposure did not directly lead to cancer but did increase tumorigenesis based on an animal experiment from Slaga et al. in which 85% of mice treated with a maximum dose of 40 mg developed carcinomas. However, benzoyl peroxide alone did not cause cancer. Researchers also observed inflammation and thickening of the skin, which suggests that longtime exposure may affect the skin (Slaga et al.).

Salicylic acid is a formulation that treats acne by removing excess oil and sebum or dead skin cells from the skin's surface. It is commonly used for comedonal acne and mild acne treatment, specifically effective at preventing clogged pores and future breakouts. Similar to benzoyl peroxide, the use of salicylic acid may lead to dryness, redness, and skin irritation (Frothingham). In a double-blind, placebo-controlled study involving 60 patients with moderate acne, a 2% salicylic acid cleanser significantly improved acne lesions after approximately eight weeks of treatment, with minimal reported side effects (Woodruff et al.). The cleanser was tolerable and did not produce excessive dryness, suggesting that salicylic acid can be incorporated into daily routines with little to no discomfort. Additional research has shown benefits beyond lesion reduction. In a 21-day study of 42 participants using a salicylic acid treatment, researchers evaluated acne severity, skin barrier, sebum production, and side effects. Results showed a 23.8% decrease in acne severity, a 23.6% decrease in excess oil or sebum production, and a 40.5% increase in skin hydration overall. Around only 5% of participants reported skin irritation or dryness, meaning salicylic acid seemed to be tolerable and improve acne (Liu et al.) (Table 1).

Azelaic acid is another synthetic ingredient commonly used in acne treatment. It helps to reduce acne by killing acnegenic bacteria, decreasing inflammation and redness, and improving post-inflammatory hyperpigmentation and acne scarring. A review of 43 randomized clinical trials found that azelaic acid was not significantly better or worse than many commonly used treatment options such as benzoyl peroxide. For example, its anti-inflammatory and

anti-microbial properties have made it a commonly prescribed medicine for acne, rosacea, and hyperpigmentation. Across sixteen acne studies, in a systematic review, azelaic acid demonstrated around a 20% reduction in acne compared to alternative topicals, like erythromycin cream, while producing limited skin reactions (King et al.) (Table 1). Similarly, Kircik reported that a 15% azelaic acid gel effectively reduced acne lesions, inflammation, and hyperpigmentation over approximately 16 weeks while being tolerable and safe. In a study involving 20 adults with mild to moderate acne, 47% of participants showed improvement by week 4, while 69% were rated as almost completely clear of acne and post-inflammatory hyperpigmentation by week 16. Furthermore, 55% of participants reported no side effects (Kircik). Although results may take longer to achieve compared to some other treatments, azelaic acid is generally associated with fewer side effects, with mild tingling and peeling being the most commonly reported.

Table 1. Selected Acne Topical Clinical Evidence

Study	Design	Comparison	Main Finding	Key Limitations
Bassett et al. (1990)	Randomized, single-blind trial; 124 participants	5% tea tree oil gel versus 5% benzoyl peroxide lotion	Both reduced inflammatory and noninflammatory lesions; benzoyl peroxide acted faster, while tea tree oil caused fewer adverse effects	Older trial; single blinding and different vehicles
Malhi et al. (2017)	Open-label, uncontrolled pilot; 18 enrolled and 14 completed	Tea tree oil gel and face wash for 12 weeks	Mean lesion counts declined over time; local peeling, dryness, and scaling were minor	No control group, small sample, and two-product regimen
Najafi-Taher et al. (2022)	Randomized clinical trial; 100 participants	Tea tree oil nanoemulsion with 0.1% adapalene versus 0.1% adapalene gel	The combination produced greater improvement than adapalene alone	Cannot isolate the effect of tea tree oil from the combination or vehicle
Hajheydari et al. (2014)	Randomized, double-blind trial; 60 participants	0.05% tretinoin plus 50% aloe vera gel versus tretinoin plus vehicle	The aloe vera combination reduced lesion scores more and produced less severe redness	Supports adjunctive use, not aloe vera monotherapy
Liu et al. (2025)	Prospective, uncontrolled study; 42 participants	Salicylic-acid-containing gel for 21 days	Acne-severity, sebum, and hydration measures improved; transient itching was uncommon	Short follow-up, no randomized comparator, mixed formulation, and industry affiliations
King et al. (2023)	Systematic review; 43 randomized trials across conditions, including 16 acne studies	Topical azelaic acid versus vehicle or active comparators	Acne studies generally favored azelaic acid over vehicle; few consistent differences appeared against active comparators	Concentrations, comparators, conditions, and outcomes varied

Discussion

The literature on medicinal and botanical plant treatments suggest that some plant-derived ingredients may offer benefits for acne management, compared to solely relying on synthetic treatment plans. Of these ingredients, the antimicrobial and anti-inflammatory properties of green tea extract and tea tree oil demonstrated the greatest potential efficacy in acne treatment (Bosslett; Parvizi et al.). However, the efficacy of medicinal plants varied depending on the specific ingredient formulation. While tea tree oil and aloe vera generally produced little to no adverse effects, their efficacy was often less consistent across studies (Bassett et al.; Hajheydari et al.) (Table 1). The limited amount of clinical trials performed for those natural ingredients makes it difficult to conclude whether natural ingredients should be used for acne management or replace current treatment plans (Yang et al.).

In comparison, synthetic ingredients, which are studied more extensively, have been found to be generally more effective than natural acne treatments, although they are also usually associated with greater health risks (Ramsis et al.). Common effects include dryness, peeling, redness, and irritation, particularly when products are used excessively or incorrectly, leading to allergic reactions and hormonal disruption (Alnuqaydan). While synthetic ingredients have shown effectiveness in reducing acne lesions, some authors have raised concerns regarding the potential health effects of certain synthetic compounds used in cosmetic and skincare products. Similarly, it was noted that overuse and exposure to synthetic materials and ingredients can lead to adverse health consequences (Wirtu). However, similar to natural ingredients, these outcomes vary depending on the ingredient and amount of exposure.

The findings of this review suggest that although both natural and synthetic over-the-counter topical ingredients can be beneficial in treating acne vulgaris, they differ in terms of safety and efficacy. While natural ingredients like tree oil showed significant reductions in acne from multiple studies like the 12-week study which reported a 54% reduction in lesions (Malhi et al.), many other studies evaluating natural ingredients involved relatively small sample sizes or tested the ingredients with synthetic medications, rather than alone (Table 1). For example, aloe vera showed greater reductions in acne lesions when combined with tretinoin than tretinoin alone (Hajheydari et al.) (Table 1). Overall, aloe vera seems to contribute more strongly to processes associated with healing and inflammation reduction, leading directly to acne lesion reduction. In contrast, benzoyl peroxide, salicylic acid, and azelaic acid reduced acne severity across multiple human trials. A 51% reduction in non-inflammatory lesions was observed when using benzoyl peroxide and niacinamide (Kaewsanit et al.), and a 23.8% decrease in acne severity was seen after only 21 days of salicylic acid treatment (Liu et al.) (Table 1). Similarly, 69% of participants using azelaic acid gel achieved almost clear skin by Week 16 (Kircik). In all, the synthetic ingredients researched utilized large sized samples, randomized controlled trials, and careful measures for the final conclusion. Even though synthetic topicals were more associated with adverse effects like dryness, peeling, redness, and itchiness, their clinical use remains highly tolerable when applied in moderation, with no evidence of severe adverse health risks (Frothingham). Additionally, some of the safety concerns may not be fully reliable due to

the way in which some of the studies were conducted and the large amount of the topical being used. Specifically, in a benzoyl peroxide study, mice were used with unusually high doses of this treatment over a 52-week period (Slaga et al.). These conditions are unusual for acne treatment in humans, therefore making the health concerns smaller than were talked about. While natural ingredients do seem promising as treatments that can help improve tolerability and reduce inflammation, the current body of literature seems limited in natural ingredients, and synthetic ingredients provide a more reliable and greater impact in terms of the efficacy in improving acne over time.

The findings have considered that combining natural and synthetic ingredients would allow for a more blended acne treatment option. The natural ingredients may help reduce irritation, while the synthetic topicals can produce stronger acne-fighting results. However, few studies have actually experimented with the use of both natural and synthetic ingredients. Future research should focus more on those clinical trials, evaluating the two types of ingredients together.

These sources together demonstrate a trade-off between efficacy and tolerability. Natural ingredients seem to be better suited for individuals with sensitive skin or those seeking acne treatments that have less side effects, while synthetic ingredients may be more suitable for individuals with more severe acne who prioritize effectiveness of treatments. Therefore, treatment selection should be based on individual skin sensitivity, skin type, and acne severity.

Acknowledgements

I would like to thank my mentor, Austin Johnson MD for supporting and guiding me through my entire process of writing this research paper.

Works Cited

Alnuqaydan, Abdullah M. "The Dark Side of Beauty: An in-Depth Analysis of the Health Hazards and Toxicological Impact of Synthetic Cosmetics and Personal Care Products." Frontiers in Public Health, vol. 12, Aug. 2024, p. 1439027. PubMed Central, https://doi.org/10.3389/fpubh.2024.1439027.

Bassett, Ingrid B., et al. "A Comparative Study of Tea‐tree Oil versus Benzoylperoxide in the Treatment of Acne." Medical Journal of Australia, vol. 153, no. 8, Oct. 1990, pp. 455–58. DOI.org (Crossref), https://doi.org/10.5694/j.1326-5377.1990.tb126150.x.

Bosslett, Marie. Botanical Therapies for Acne: A Clinical Review of Social Media Trends. 7 Jan. 2026, https://www.dermatologytimes.com/view/botanical-therapies-for-acne-a-clinical-review-of-social-media-trends.

Draelos, Zoe, et al. "A 12‐Week Randomized, Double‐Blind, Placebo‐Controlled Trial for the Efficacy and Safety of a Novel Nutraceutical for Mild‐to‐Moderate Acne." Journal of Cosmetic Dermatology, vol. 24, no. 5, May 2025, p. e70220. PubMed Central, https://doi.org/10.1111/jocd.70220.

Frothingham, Scott. "Salicylic Acid for Acne: Benefits, Dosages, and Side Effects." Healthline, 12 Dec. 2019, https://www.healthline.com/health/skin/salicylic-acid-for-acne.

Hajheydari, Zohreh, et al. "Effect of Aloe Vera Topical Gel Combined with Tretinoin in Treatment of Mild and Moderate Acne Vulgaris: A Randomized, Double-Blind, Prospective Trial." The Journal of Dermatological Treatment, vol. 25, no. 2, Apr. 2014, pp. 123–29. PubMed, https://doi.org/10.3109/09546634.2013.768328.

Institute, Dermatology. "Are Natural Remedies for Acne as Effective." Dermatology Institute, 6 Apr. 2023, https://dermsurgery.net/are-natural-remedies-for-acne-as-effective-as-prescription-options

Jisa, Tajin Ahmed, et al. "Prevalence of Acne and Its Impact on Quality of Life, Social Appearance Anxiety and Treatment Practices Among Young Adults." Journal of Cosmetic Dermatology, vol. 25, no. 1, Jan. 2026, p. e70654. PubMed Central, https://doi.org/10.1111/jocd.70654.

Kaewsanit, Tarnyamas, et al. "Clinical Comparison of Topical 2.5% Benzoyl Peroxide plus 5% Niacinamide to 2.5% Benzoyl Peroxide Alone in the Treatment of Mild to Moderate Facial Acne Vulgaris." Journal of Clinical Aesthetic Dermatology, 2021, https://jcadonline.com/comparison-benzoyl-peroxide-niacinamide-acne/.

Kairey, Lana, et al. "Efficacy and Safety of Melaleuca Alternifolia (Tea Tree) Oil for Human Health—A Systematic Review of Randomized Controlled Trials." Frontiers in Pharmacology, vol. 14, Mar. 2023. Frontiers, https://doi.org/10.3389/fphar.2023.1116077.

King, Sarah, et al. "A Systematic Review to Evaluate the Efficacy of Azelaic Acid in the Management of Acne, Rosacea, Melasma and Skin Aging." Journal of Cosmetic Dermatology, vol. 22, no. 10, Oct. 2023, pp. 2650–62. DOI.org (Crossref), https://doi.org/10.1111/jocd.15923.

Kircik, Leon H. “Efficacy and Safety of Azelaic Acid (AzA) Gel 15% in the Treatment of Post-Inflammatory Hyperpigmentation and Acne: A 16-Week, Baseline-Controlled Study.” JDDonline - Journal of Drugs in Dermatology, June 2011, https://jddonline.com/articles/efficacy-and-safety-of-azelaic-acid-aza-gel-15-in-the-treatment-of-post-inflammatory-hyperpigmentati-S1545961611P0586X/.

Liu, Ye, et al. “Clinical Efficacy of a Salicylic Acid–Containing Gel on Acne Management and Skin Barrier Function: A 21‐Day Prospective Study.” Journal of Cosmetic Dermatology, vol. 24, no. 7, July 2025, p. e70353. PubMed Central, https://doi.org/10.1111/jocd.70353.

Malhi, Harsimran Kaur, et al. “Tea Tree Oil Gel for Mild to Moderate Acne; a 12 Week Uncontrolled, Open-Label Phase II Pilot Study.” The Australasian Journal of Dermatology, vol. 58, no. 3, Aug. 2017, pp. 205–10. PubMed, https://doi.org/10.1111/ajd.12465.

Mayo Clinic Staff. “Benzoyl Peroxide (Topical Route) - Side Effects & Dosage.” Mayo Clinic, https://www.mayoclinic.org/drugs-supplements/benzoyl-peroxide-topical-route/description/drg-20062425.

“Tea Tree Oil.” Mayo Clinic, 15 Aug. 2025, https://www.mayoclinic.org/drugs-supplements-tea-tree-oil/art-20364246.

Mayo Clinic Staff. “Tea Tree Oil.” Mayo Clinic, 15 Aug. 2025, https://www.mayoclinic.org/drugs-supplements-tea-tree-oil/art-20364246.

Najafi-Taher, Roqya, et al. “A Topical Gel of Tea Tree Oil Nanoemulsion Containing Adapalene versus Adapalene Marketed Gel in Patients with Acne Vulgaris: A Randomized Clinical Trial.” Archives of Dermatological Research, vol. 314, no. 7, Sept. 2022, pp. 673–79. Springer Link, https://doi.org/10.1007/s00403-021-02267-2.

Nascimento, Tânia, et al. “Tea Tree Oil: Properties and the Therapeutic Approach to Acne—A Review.” Antioxidants, vol. 12, no. 6, June 2023, p. 1264. PubMed Central, https://doi.org/10.3390/antiox12061264.

Pal, Suraj, et al. “The Effect of Aloe Vera on Skin and Its Commensals: Contribution of Acemannan in Curing Acne Caused by Propionibacterium Acnes.” Microorganisms, vol. 12, no. 10, Oct. 2024, p. 2070. PubMed Central, https://doi.org/10.3390/microorganisms12102070.

Parvizi, Mohammad Mahdi, et al. “Medicinal Plants for Acne Vulgaris: An Evidence‐Based Review of Treatments Promoted by Social Media.” Journal of Cosmetic Dermatology, vol. 25, no. 1, Jan. 2026, p. e70628. PubMed Central, https://doi.org/10.1111/jocd.70628.

Ramsis, Triveena, et al. “The Role of Current Synthetic and Possible Plant and Marine Phytochemical Compounds in the Treatment of Acne.” RSC Advances, vol. 14, no. 33, 2024, pp. 24287–321. pubs.rsc.org, https://doi.org/10.1039/D4RA03865G.

Reynolds, Rachel V., et al. “Guidelines of Care for the Management of Acne Vulgaris.” Journal of the American Academy of Dermatology, vol. 90, no. 5, May 2024, p. 1006.e1-1006.e30. Mayo Clinic, https://doi.org/10.1016/j.jaad.2023.12.017.

Slaga et al., T. J. “Skin Tumor-Promoting Activity of Benzoyl Peroxide, a Widely Used Free Radical-Generating Compoun.” Science, vol. 213, no. 4511, 1981, pp. 1023–25.

Sutaria, Amita H., et al. “Acne Vulgaris.” StatPearls [Internet], StatPearls Publishing, 2023. www.ncbi.nlm.nih.gov, https://www.ncbi.nlm.nih.gov/sites/books/NBK459173/.

Torjesen, Ingrid. “Natural Extracts Outweigh Synthetic Antibiotics for Acne Treatment.” Dermatology Times, vol. 40, no. 3, 14 Jan. 2019, https://www.dermatologytimes.com/view/natural-extracts-outweigh-synthetic-antibiotics-acne-treatment.

Wanner, Molly, and Neera Nathan. “Clean Cosmetics: The Science behind the Trend.” Harvard Health, 4 Mar. 2019, https://www.health.harvard.edu/blog/clean-cosmetics-the-science-behind-the-trend-20190 30416066.

Wirtu, Yohannes Desalegn. “A Review of Environmental and Health Effects of Synthetic Cosmetics.” Frontiers in Environmental Science, vol. 12, Oct. 2024. Frontiers, https://doi.org/10.3389/fenvs.2024.1402893.

Woodruff, Jaqueline, et al. “A Double-Blind, Placebo-Controlled Evaluation of a 2% Salicylic Acid Cleanser for Improvement of Acne Vulgaris.” Journal of the American Academy of Dermatology, vol. 68, no. 4, Apr. 2013, ScienceDirect. NLM, https://doi.org/10.1016/j.jaad.2012.12.052.

Yang, Ji Hoon, et al. “Seeking New Acne Treatment from Natural Products, Devices and Synthetic Drug Discovery.” Dermato-Endocrinology, vol. 9, no. 1, Oct. 2017, p. e1356520. PubMed Central, https://doi.org/10.1080/19381980.2017.1356520.

Zhong, Hongyu, et al. “Efficacy of a New Non-Drug Acne Therapy: Aloe Vera Gel Combined With Ultrasound and Soft Mask for the Treatment of Mild to Severe Facial Acne.” Frontiers in Medicine, vol. 8, May 2021, p. 662640. PubMed Central, https://doi.org/10.3389/fmed.2021.662640.

The Effects of Camber and Thickness on Aerodynamic Performance of NACA Four-Digit Airfoils: A Simulation Study with Experimental Flow Visualization By Landry Silvers

Abstract

Airfoils are fundamental shapes to aircraft, racing cars, wind turbines and industrial turbomachinery. Their geometry determines the aerodynamic forces of lift and drag, as well as stability. The two most influential characteristics of an airfoil are its camber, or the maximum displacement of the mean line from the chord line, and its maximum thickness to chord ratio. This paper evaluates the effects of geometric thickness and camber on lift, drag, L/D ratios and overall flight performance for five four-digit airfoils: NACA 0009, 0012, 0015, 2012 and 4012. The objective was to identify and discuss significant aerodynamic differences produced by minor geometry variations at a relatively low Reynolds number, while comparing the results to a student-friendly educational simulator. The results of calculations performed using NASA's FoilSim Student Interactive Edition at angles of attack varying from -20° to +20° with 2° steps at the Reynolds number of 1.3×10^4 were compared to the results of a physical experiment performed in a custom-built wind tunnel. The five airfoils were subjected to the same test in the physical tunnel at angles of attack of -31.4°, -11.8°, 0°, +11.8° and +31.4°, while using laser-fog visualization. FoilSim's lift-curve slopes were approximately 0.111-0.117 per degree, which is within 7% of the thin-airfoil value of 0.110, and each percent of camber shifted the zero-lift angle by about -1.0°. Increasing the airfoil thickness from 9% (NACA 0009) to 15% (NACA 0015) raised its zero-degree drag coefficient from 0.022 to 0.041 and decreased the peak L/D ratio from 14.0 to 12.3. On the other hand, camber increased the maximum coefficient of lift from 1.40 (NACA 0012) to 1.73 (NACA 4012) and shifted the peak lift-to-drag ratio to 0° angle of attack, where NACA 4012 achieved a value of 13.8. Compared to experimental data found in the literature, FoilSim's drag coefficients are justifiable at this Reynolds number, but its maximum lift coefficients are not. The objective of the physical model was to provide a visual confirmation of the trends observed by FoilSim calculations: flow turning at zero degrees grew visibly with camber, and every section separated fully at ±31.4°. The comparison also served to highlight two issues with the educational software: asymmetric drag coefficients for positive and negative angles of attack for symmetric airfoils and constant lift coefficients for varying Reynolds numbers.

1. Introduction

Airfoils are the simplest cross-section shapes from which all airplanes, gliders, or spacecraft wings, as well as control surfaces, are made. They are responsible for the aircraft's aerodynamic characteristics, lift, drag, and stability. However, among the parameters describing the geometry, the camber, or maximum displacement of the mean line from the chord line, and the maximum thickness to chord ratio are two of the most significant.

The thickness distribution is crucial for the strength of the wing and pressure recovery on the wing's suction side, which influences the drag. Meanwhile, the camber determines the lift characteristics at zero angle of attack. All subsonic aircraft wings, propellers, and wind-turbine blades are compromises between the two parameters, which are usually optimized for each particular aircraft, considering the scale. For example, for small-scale models, laminar flow separation occurs more frequently than for full-size airplanes.

The NACA four-digit series is one of the most reliable and popular series created in the 1930s and described in detail in the works of Abbott and von Doenhoff (124). The main advantage of the four-digit series is encoding of the parameters described above in the name of the profiles. Therefore, it becomes easy to perform comparative analysis of lift, drag, and lift-to-drag ratios of profiles with different camber and maximum thickness to chord ratio. This study uses the simplicity of the four-digit naming system to explore the dependence of aerodynamics on thickness and camber. The family of thickness series (NACA 0009, NACA 0012, NACA 0015) keeps the camber at zero level and changes the thickness from 9% to 15%, while the family of camber series (NACA 0012, NACA 2012, NACA 4012) keeps the thickness at 12% and changes the camber from 0% to 4% (Fig. 2). The NACA 0012 profile is used as the reference in both series.

The aerodynamic characteristics of the profiles listed above are well-established with high confidence levels from nearly a century of research. However, the aim of this study is to analyze the tool which provides basic aerodynamic knowledge to students, visualizes the airflow, and estimates lift, drag, and lift-to-drag ratio for any given profile. NASA FoilSim is probably the most popular tool among educators which provides reasonable estimates for all three parameters for any given airfoil with a certain angle of attack. However, the accuracy of FoilSim is rarely questioned by its users.

This paper aims to analyze FoilSim performance for this particular set of airfoils. First of all, the FoilSim estimations of lift, drag, and lift-to-drag ratio for these airfoils were compared to the thin-airfoil theory and experimental data retrieved from literature. Additionally, the flow patterns visualized in FoilSim were compared to the flow structures obtained experimentally in the wind tunnel test. The paper aims at three research questions.
First, how do the changes in thickness and camber influence the lift, drag, and lift-to-drag ratio according to FoilSim? Secondly, what changes in flow structures can be expected for the changes in thickness and camber analyzed? Finally, how accurate are the estimations provided by FoilSim?

The expected results rely on the thin-airfoil theory. According to the theory, increasing the thickness will increase the drag at zero angle of attack, while keeping the lift coefficient at zero angle of attack unchanged. Additionally, increasing the thickness will decrease the maximum lift-to-drag ratio without changing the lift slope. On the other hand, increasing the camber will shift the lift curve to the left by approximately one degree per one percent of camber, while keeping the lift slope constant. These are two separate expectations which can be validated with the use of FoilSim.

2. Aerodynamic Theory

2.1 The four-digit convention

A NACA four-digit airfoil designation gives the maximum camber as a percentage of the chord (first digit), location of the maximum camber along the chord (second digit) and the maximum thickness as a percentage of the chord (last two digits). The NACA 2412 profile, for instance, has 2% camber at 40% of the chord and 12% thickness. Fig. 1 shows the location of the camber and thickness. The airfoil sections in this study are designated by the camber and thickness values used in the FoilSim input forms. Exceptions to this convention occur because the FoilSim airfoil module generates shapes using a conformal transformation of a Joukowski profile and the designation is only approximate to the actual NACA profiles used in the wind tunnel tests (see Section 3.3). The convention is nevertheless retained for simplicity in referring to the different sections. Note that a 00-series designation corresponds to symmetric profiles which have zero lift at zero angle of attack.

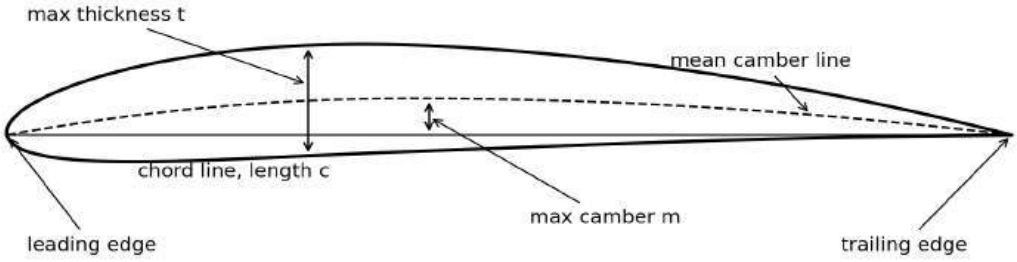


Fig 1: Geometric parameters of a NACA four-digit section: chord line, mean camber line, maximum camber m, and maximum thickness t. Drawn from the standard NACA equations with 4% camber at 0.4c and 12% thickness; FoilSim's Joukowski-derived shapes approximate this geometry (Section 2.1).

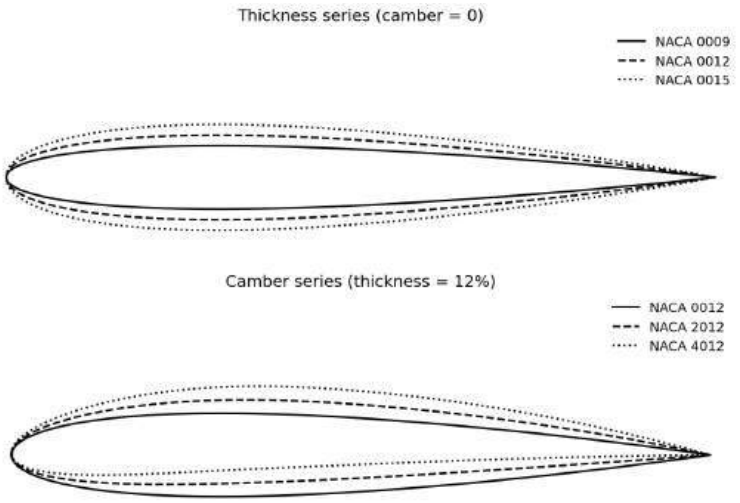


Fig 2: The five sections studied, drawn to a common chord. Top: thickness series (NACA 0009, 0012, 0015). Bottom: camber series (NACA 0012, 2012, 4012). Camber position as in Fig. 1.

2.2 Lift and the effect of camber

An airfoil gains lift if the pressure on its upper surface is, on average, lower than the pressure on its lower surface. This difference in pressure can be related to circulation: the flow must leave the trailing edge smoothly (the Kutta condition), and this requirement fixes a unique value of circulation for a given section and angle of attack, related to lift per unit span via the Kutta-Joukowski theorem (Anderson 267). Thin-airfoil theory simplifies this aerodynamic complexity to two findings relevant to this study. First, the lift coefficient increases linearly with angle of attack at a rate of 2π per radian, approximately 0.110 per degree, independent of the section profile. Second, the camber does not affect the slope of the lift curve but shifts it upwards, allowing positively cambered sections to produce positive lift even at zero incidence and reach zero lift at a negative angle of attack, approximately one degree per percent camber for the families of sections studied.

2.3 Drag and the effect of thickness

Profile drag at these speeds is the sum of skin friction and pressure drag. Thickness will affect the second term. In particular, it will raise it by causing the flow to accelerate more strongly over the front half of the section (so that it has to decelerate more strongly over the rear half) and, therefore, by creating a more adverse pressure gradient over the rear half. This chain thickens the boundary layer before separation, and the separated region creates a deficit of pressure on the aft-facing surfaces of the profile, inducing additional form drag. Hence, the minimum drag is higher at greater thickness, and the corresponding lift-to-drag ratios are lower.

2.4 Boundary layers and Reynolds number

How much of this margin exists depends on the Reynolds number, $Re = \rho Vc/\mu$, the ratio of inertial to viscous forces based on chord length, c, and speed, V . At $Re \approx 10^6$, boundary layers transitioning to turbulence early and resisting separation dominate, while at $Re \approx 10^4$, the realm of small-scale models in benchtop tunnels, boundary layers are laminar, detach easily under even gentle adverse pressure gradients, and may support laminar separation bubbles. The same wing then has poorer performance at model scale than suggested by high-Re bench tests, stalling earlier and generating more drag. This is an important consideration in the present case, as the simulation was performed at a condition close to the tunnel, namely $Re \approx 1.3\times10^4$ in FoilSim compared to $Re \approx 1.0\times10^4$ in the test section, as computed from the measured chord (109.4 mm) and speed (1.4 ± 0.3 m/s; Section 3.2). Published data for the relevant test sections, by contrast, fall around $Re = 3\times10^6$ and the comparison to the literature thus spans the gap, and is handled in detail in the Discussion.

2.5 Stall and efficiency

As angle of attack increases, lift increases linearly with angle until the suction-side boundary layer can no longer be held on the airfoil by the pressure gradient, whereupon lift decreases and drag increases sharply (stall). The lift-to-drag ratio, which is a commonly accepted

figure of merit for section efficiency, peaks at a moderate angle of attack before the stall, where lift is still substantial but the penalty in drag due to angle of attack is not yet prohibitive. The location of this peak, and its value, are used as a basis for comparison in the present study.

2.6 What a simplified model can capture

FoilSim is not a Navier-Stokes solver. It calculates an ideal lift from an ideal potential flow plus some empiricism, and an approximate drag from stored values, rather than from a numerical solution of the boundary layer. A model of this type should be able to reproduce results which are determined by geometry and circulation, for instance, the lift-curve slope, or the zero-lift angle due to camber, but would fail to predict forces with a dependence on the boundary layer, like absolute drag, or a difference in drag between two profiles at the same lift coefficient. Part of the reason for this work was to test this expectation.

3. Methodology

3.1 Computational procedure

Aerodynamic coefficients were computed in NASA FoilSim Student Interactive Edition (NASA Glenn Research Center) for the five sections at angles of attack from −20° to +20° with a 2° step. The simulations used the same flow conditions as the wind tunnel tests: the airspeed was set to 3.5 mph (1.57 m/s), chord to 0.400 ft (the closest available to the 0.359 ft model chord), span to the minimum 5.000 ft available in the software, and standard sea level air. According to the software, this corresponds to a chord Reynolds number of 13,489, which is about 34% above the 1.0×10^4 value of the wind tunnel, a difference of the same order as the ±20% uncertainty in the measured test-section speed (Section 3.2). Because the span is minimum in the software, the aspect ratio is fixed at 12.5; this means that FoilSim's finite-wing corrections decrease all coefficients compared to their 2D values (the correction is applied to all sections alike, so it cancels out when comparing lift-curve slopes, but not the drag coefficients, which depend on lift). For each case, the lift coefficient, drag coefficient, and lift-to-drag ratio were recorded on a spreadsheet. The summary values are presented in Table 1; the entire set of 105 calculations is provided with this paper as an Excel attachment.

3.2 Wind tunnel

The wind tunnel was 3D printed from the open-source DIY design published on Printables by the user JoDon3311 under the name Open Return Wind Tunnel (JoDon3311). This design prints in roughly 1500 g of PLA, costing around 60 USD in filament (Fig. 3a). The published design only includes the printed parts; the fan (a 120 mm computer fan), the two flow-conditioning grids, the laser and its mount, and the fog generation, settling, and seeding system were all selected, designed, or built by the author for this project. The tunnel is a suction-type open circuit: air enters at the upstream end, is accelerated through a printed contraction, passes through the test section bounded by acrylic sheets, and is drawn through the

diffuser and exhausted by the fan at the downstream end. The flow is conditioned upon entry with a grid, reducing turbulence, and ensuring consistent flow at the test section inlet. Speed was controlled by a variable DC power supply (3–12 V) connected to the fan motor. The power supply was set to the same value (approximately 60 % of the DC 3-12 V range) for all the trials presented in this paper; this allowed direct comparison between the results and ensured that the same flow regime was used for all cases. As shown in the videos included as supporting information, test-section flow velocity was assessed by measuring the advection of water fog: the fog filaments' trajectories are seen as sloped lines in the plots generated by arranging horizontal bands of consecutive video frames as pixels, with the 150.0 mm window width providing the scale. An independently calculated value agrees: according to the manufacturer's specs, the Easy Cloud 120×25 mm fan (i.e., the 120 mm diameter impeller) develops 95.8 CFM at 2800 RPM and 1.66 m^3/min at the 60% reduction as was used in all tests described in this paper; hence, the free-air velocity would be 1.1 m/s at the fan inlet for the 182.4 × 137.2 mm test-section cross-section, which is within the 1.4 ±0.3 m/s range found experimentally using the streak method. The test section is 182.4 mm (7.181 in) wide, 137.2 mm (5.4 in) tall, and 245.0 mm (9.65 in) long between the end plates, with a 150.0 × 137.2 mm view window. The latter was cut from a transparency sheet of an office chair mat with an exacto knife; the view is limited to low resolution due to the material's quality and the cutting precision, and the resulting surface imperfections are visible in some of the images in Section 4.

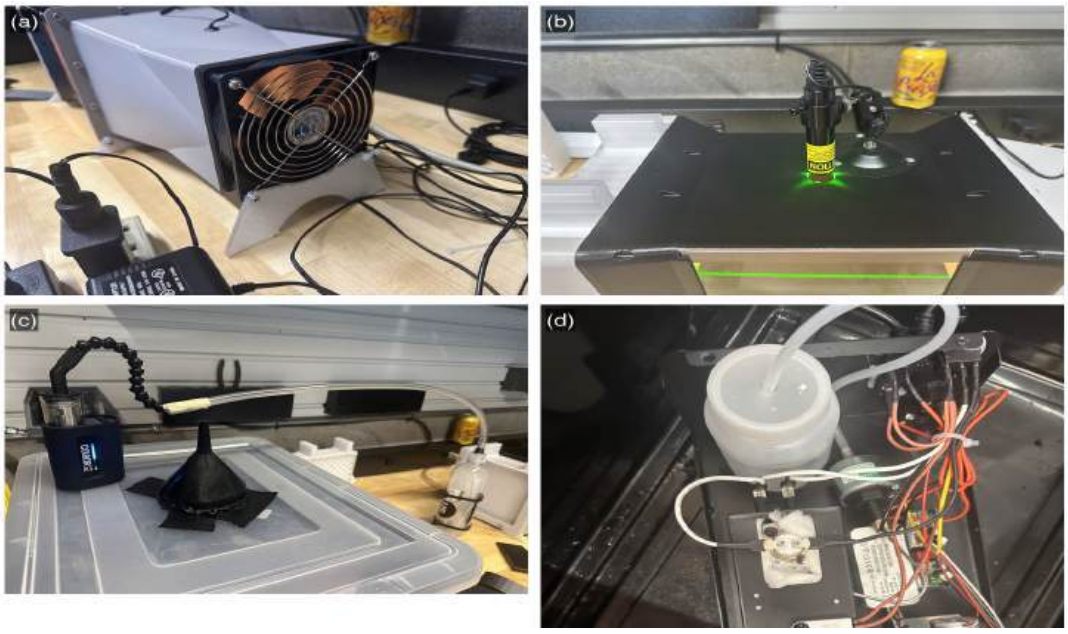


Fig 3: Apparatus. (a) Tunnel ductwork printed from the open-source DIY Open Return Wind Tunnel design (JoDon3311, Printables); fan, grids, and all instrumentation added by the author. (b) Line laser on an articulated mount, designed for this study, projecting a vertical light sheet at model midspan. (c) Final fog-seeding system: the fog machine discharges into a settling container that feeds a perforated water-bottle manifold at the tunnel inlet. (d) First-generation fog enclosure, shown disassembled: the machine's heating element was sealed inside the box and began melting it; the design was abandoned for (c).

Two interchangeable flow-conditioning grids were designed and printed for this study: a fine lattice (characteristic cell pitch of approximately 4.5 mm and bar width of approximately 0.9 mm) and a coarse lattice (characteristic cell pitch of approximately 14 mm and bar width of approximately 1.7 mm) (Fig. 4). The main experiments were carried out with the fine grid, whereas the NACA 0012 profile was re-investigated with the coarse grid in order to estimate the influence of the inlet flow conditioning on the visualization results (see Section 4.3).

Fig 4: The two 3D-printed flow-conditioning grids: coarse lattice (left, ≈13 cells across, pitch ≈ 14 mm) and fine lattice (right, ≈40 cells across, pitch ≈ 4.5 mm). Dimensions estimated photogrammetrically.

Flow was seeded with theatrical fog. Initially, to contain the fog, the seeding system was constructed in two iterations. The first placed the fog machine, including its heating element, in a plastic enclosure (Fig. 3d), which the heat from the latter quickly melted, prompting the search for an alternative. In the final configuration, the machine feeds fog directly into a constricted settling volume from which it exits through a perforated water-bottle manifold at the tunnel's inlet. The manifold's perforations damp the pulsed jet exiting the machine, choking it to a low-speed flow that feeds the test section at the model's mid-span (Fig. 3c). A green line laser projects a vertical sheet of light intersecting the freestream at model mid span, allowing visualization of the flow as streaklines of suspended fog. The room was darkened for all tests.

3.3 Airfoil models

Five profiles were extruded in Onshape CAD to a span of 50.8 mm (2 in), a chord of 109.4 mm (0.6 of 182.4 mm), and a custom mounting hole to fit the tunnel's angle-of-attack

fixture. They were printed in PLA (preprocessed in PrusaSlicer) and tested as-printed; no additional texture was added to their surfaces, which are otherwise smooth. The models are each 28% of the test section width and have an aspect ratio of 0.46, a value insufficient to produce a purely 2D flow, meaning that the tip vortices affect the flow even at mid-span, where the laser sheet is placed. The models were mounted in the test section via 3D-printed pedestals at a particular angle of attack, selected among 0°, ±11.8°, and ±31.4°. The nonzero values were defined by the geometry of the three-dimensional printed model in the figure above relative to the tunnel schematics, with an uncertainty of the order of ±0.5°, or approximately the resolution of the protractor used to set the angle during the experiment. The nonzero angles bracket the expected stall angle for all airfoils tested.

3.4 Image acquisition

Each section-view combination was photographed through the test-section window using an iPhone 14 Pro mounted on a tripod (4K, 24 fps for the videos; for the still images, the phone was at the same position and viewing angles for all runs at a given section, so that images can be directly compared). The specific angles were identified by cross-referencing the chord orientation in the image with the horizontal window frame; in all cases, the phone was correctly positioned according to the intended mount (level = 0 , moderate = ±11.8 , extreme = ±31.4). The images have been adjusted for contrast only, and the specific adjustments are identified in the captions.

3.5 Scope of the experimental data

The wind tunnel used for the experiment provides no force balance, so that the only experimental information available is qualitative: streakline geometry, attachment and separation characteristics, wake structure, and flow turning. All quantitative statements in the present work are based on the computational data and/or on the literature; thus, the present visualization can only be used to verify that these flow features postulated in the FoilSim indeed can occur on these geometries. The laser sheet is arranged to visualize the suction side, while the model itself shades the pressure side from view, and the support post influences the flow below the model; hence, the discussion is limited to the upper surface flow structure and wake characteristics.

4. Results

Computational results are presented first (Section 4.1), then the flow visualization (Section 4.2) and the inlet-conditioning comparison (Section 4.3).

4.1 Computational results

Fig. 5 and 6 present the computed lift, drag, and efficiency curves for the thickness and camber series; Table 1 collects the key metrics.

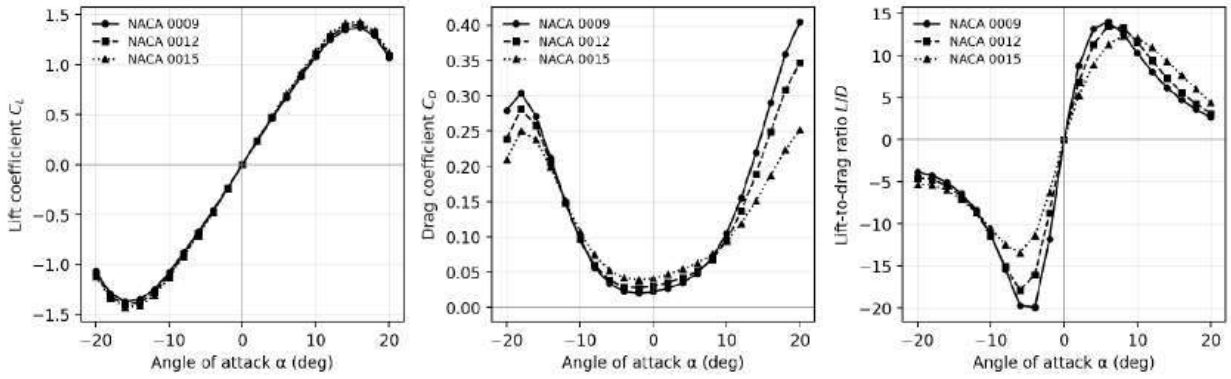


Fig 5: Computed aerodynamic curves for the thickness series (NACA 0009, 0012, 0015) at Re = 1.3×10^4: (a) lift coefficient, (b) drag coefficient, (c) lift-to-drag ratio versus angle of attack.

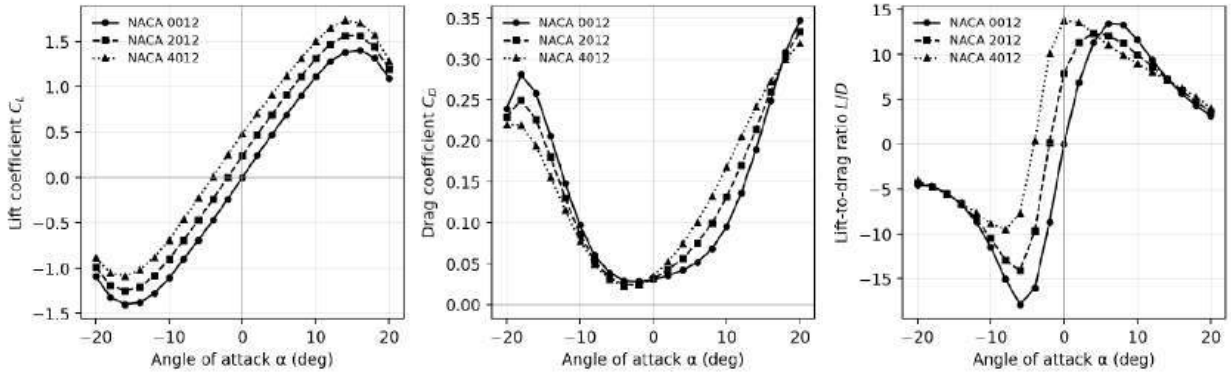


Fig 6: Computed aerodynamic curves for the camber series (NACA 0012, 2012, 4012) at Re = 1.3×10^4: (a) lift coefficient, (b) drag coefficient, (c) lift-to-drag ratio versus angle of attack.

Table 1: Summary of computed aerodynamic metrics. Lift-curve slope fitted over −8° to +8°; (L/D)max taken over positive angles.

Section	**dCL/dα (per °)**	**α_0L (°)**	**CL,max**	**α at CL,max (°)**	**CD at 0°**	**(L/D)max**	**α at (L/D)max (°)**
NACA 0009	0.111	0.0	1.37	16	0.022	14.0	6
NACA 0012	0.114	0.0	1.40	16	0.030	13.4	6
NACA 0015	0.117	0.0	1.43	16	0.041	12.3	8
NACA 2012	0.114	−2.0	1.56	14	0.031	12.3	4
NACA 4012	0.112	−4.1	1.73	14	0.035	13.8	0

Lift behaves the way thin-airfoil theory dictates: as a linear function of α for all five sections, within ±8°, having slopes from 0.111 to 0.117 per degree compared to the predicted 0.110, and a slight thickness dependence. In addition, the camber simply shifts the line up and down: the slope remains nearly the same, while the zero-lift angle α_0L changes by nearly 1° per

percent camber, becoming -2.0° for the 2012 and -4.1° for the 4012: almost exactly the one degree per percent camber predicted by the theory. However, this is true only within the linear range; beyond the stall, the agreement is not so good. The model predicts that the cambered sections have a higher maximum lift coefficient, 1.56 and 1.73 vs. 1.40, but at a lower angle of attack (14° vs. 16°), so that camber in FoilSim is not merely a shift of the coordinate system once the flow stops being attached.

At zero incidence, CD increases roughly by a factor of two over the series from the thinnest 0009 to the 0012 to the 0015: 0.022, 0.030, and 0.041 respectively, with virtually the same spread at each angle up to +8°; beyond that the order is reversed, and by +20° the 0015 has the lowest drag of the three. Furthermore, it is worth noting that the drag calculation itself has an inconsistency embedded in it, namely that for a symmetric profile the flow at −α is the exact opposite of the flow at +α, implying CD (−α)=CD (+α) by symmetry, which is not the case for any of the three symmetric profiles tested. The three symmetric sections have their minimum drag at −2° incidence rather than zero, and negative angles of attack, up to the point of stall, have lower drag than their positive counterparts, by as much as 34% at ±4° and 0.014 at ±6°. The sign of the discrepancy is not even consistent across the thickness series: between ±10° and ±18° the 0012 and the 0015 have higher drag at negative angles, the 0015 by 24% at ±14°, before reverting to the opposite ordering at ±20° (see Fig. 5b). Similarly notable is the approximate antisymmetry, within calculation accuracy, of the lift coefficients reported by the program: for instance, in the last case just quoted, CL (−20°)=− CL (+20°). Obviously, the lift and drag calculations utilize fundamentally different algorithms; Section 5 has a few words to say on this matter.

The efficiency of the profile, its lift-to-drag ratio, peaks at increasingly higher angles of attack with increasing thickness, and shifts toward negative α with camber, as expected: thus, the 0012 peaks at 6°, the 2012 at 4°, whereas the 4012 reaches a maximum of L/D = 13.8 at zero incidence. That this shift constitutes the practical value of camber is evident from the discussion in Section 5. Meanwhile, the large negative values of L/D in Figs 5c and 6c (reaching a minimum of −20 for the 0009 at −4° incidence) are a direct result of the asymmetry of the drag calculation mentioned above, and are discussed no further here.

Two brief comments are due before proceeding: first, that the program stalls every profile at around 14–16° and then gradually recovers, its lift coefficient peaking gently and symmetrically around the stall angle within the range of ±20°, whereas in reality separation occurs at much lower angles of attack for these profiles, around $Re = 10^4$, and is discussed in detail in subsection 4.2. Secondly, as a check of how the calculations respond to Reynolds number, the same computations were performed at $Re = 7.6\times10^5$: the lift coefficients agreed with those of the $Re = 1.3\times10^4$ runs at every angle to the precision of the last decimal place, whereas the drag coefficients differed by approximately 10–20%. The lift coefficients computed by FoilSim are thus invariant with Reynolds number, matching to the last decimal place across a 58-fold change, while the drag coefficients are not.

4.2 Flow visualization

The photos are to be read as streakline recordings: marks are created by seeded fluid particles traveling along their pathlines. Only when the flow is steady can streaklines be identified with streamlines. The flow is interpreted only on the illuminated top surface and in the wake (see Section 3.5).

Fig. 7 shows the NACA 0012 airfoil at different angles of attack. At 0° (case a) the streaklines are straight and parallel to the section, converging to a narrow wake at the trailing edge. This indicates an attached flow, associated with zero lift. At +11.8° (case b) the bundle on the suction side widens towards the trailing edge and begins to lift off the surface, showing signs of separation typical for a positive angle of attack. The flow at -- 11.8° (case c) is the same condition as case (b) viewed from the other side, so the illuminated surface is now the pressure surface. The streaklines follow the surface from the leading edge to the trailing one without sign of separation. As the angle reaches +31.4° (case d), the suction side is completely separated. The fog is rolled up in the form of a large bubble located at the trailing edge. Similar to case (d), the flow at –31.4° (case e) is fully separated, but occurs on the pressure surface. There are no visible disturbances on the upper surface, while the wake is as wide as in case (d).

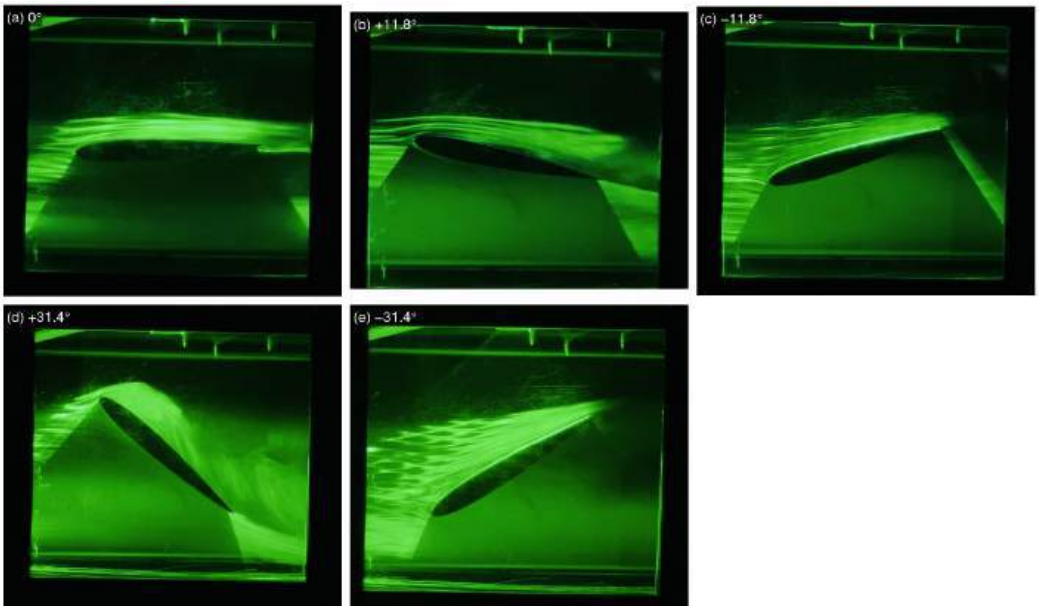


Fig 7: Streakline visualization of the NACA 0012 at the five set angles of attack. (a) 0°, (b) +11.8°, (c) −11.8°, (d) +31.4°, (e) −31.4°. Flow is left to right; the laser sheet enters from above, so only the upper surface is illuminated. On a symmetric section the ±11.8° pair is aerodynamically one condition viewed from both sides: (b) shows the suction surface, already partially separated at this Reynolds number, while (c) shows the attached pressure surface.

Fig. 8 presents the results which compare the camber of the flow at similar settings. For the 0012 section, none of the streaklines were deflected significantly. For the 2012, the streaklines are notably turned down past the trailing edge, while for the 4012, the turn is more pronounced, and it is already visible upstream of the model, as the incoming streaklines already

turn towards the suction surface. The turning down of the flow is the result of momentum transfer to the airfoil section, and thus these pictures are the analogs of Fig. 7, presenting lift at zero angle of attack (CL = 0 , 0.24, 0.48), and they also follow the same ordering.

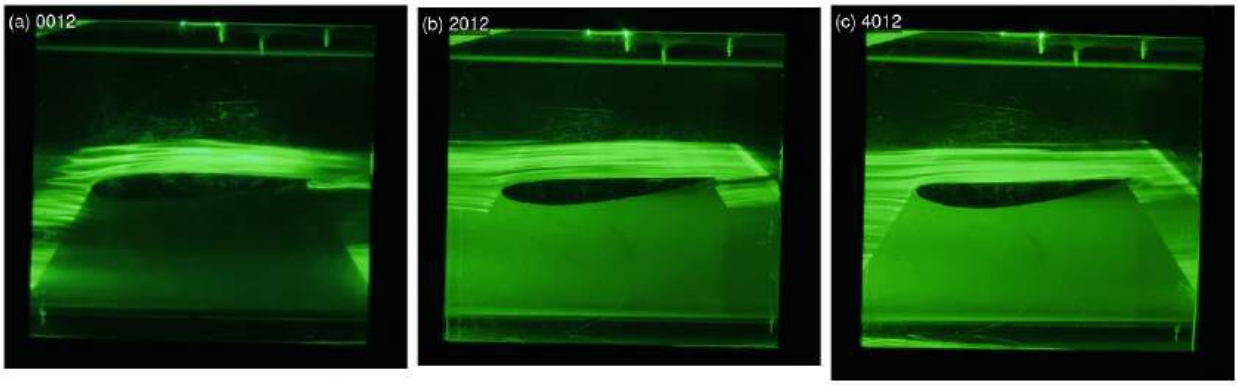


Fig 8: Effect of camber at zero geometric incidence: (a) NACA 0012, (b) NACA 2012, (c) NACA 4012. All three photographs at $\alpha = 0°$ and the same fan setting.

Fig. 9 shows the same comparison, but for thickness, and the conclusion is also that there is little effect. Attached flow on all three is seen in the streaklines, with only modest expansion and contraction around the thickest parts, and wakes that are not very different from each other given the resolution of the images. This is entirely consistent with the computed force coefficients (Section 4.1). At zero incidence, thickness affects drag substantially (0.022 versus 0.041), but this spread of 0.019 is comparable to the 0.023 by which drag rises on a single section between 0° and 6°. In particular, these differences are too small to cause major differences in streaklines, and the two methods are measuring somewhat different effects, and so both are needed.

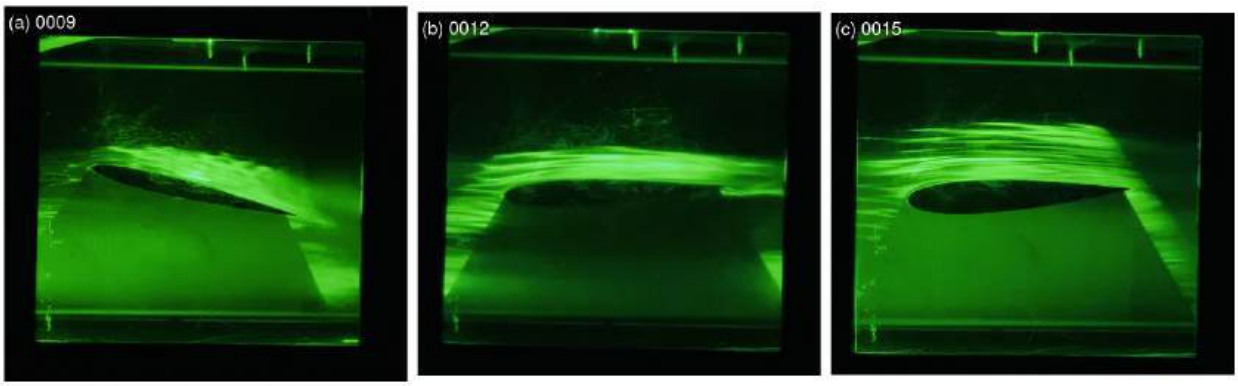


Fig 9: Thickness series: (a) NACA 0009, (b) NACA 0012, (c) NACA 0015. All panels at $\alpha = 0°$.

Fig. 10 shows the deep-stall conditions at +31.4 degrees. All three airfoils have fully separated flows, with the shear layer rolling back from the leading edge and a broad wake region, but on the cambered profiles, most notably the 4012, the mist is organized into a leading-edge

vortex that sheds downstream as well. The camber appears to modify the separated-flow structure but not eliminate separation, as suggested by the comparison of the computed lift curves: at 20 degrees angle of attack, all three have roughly comparable lift, less than their respective maximums, and beyond ±20 degrees, the model has nothing to compare to these photos.

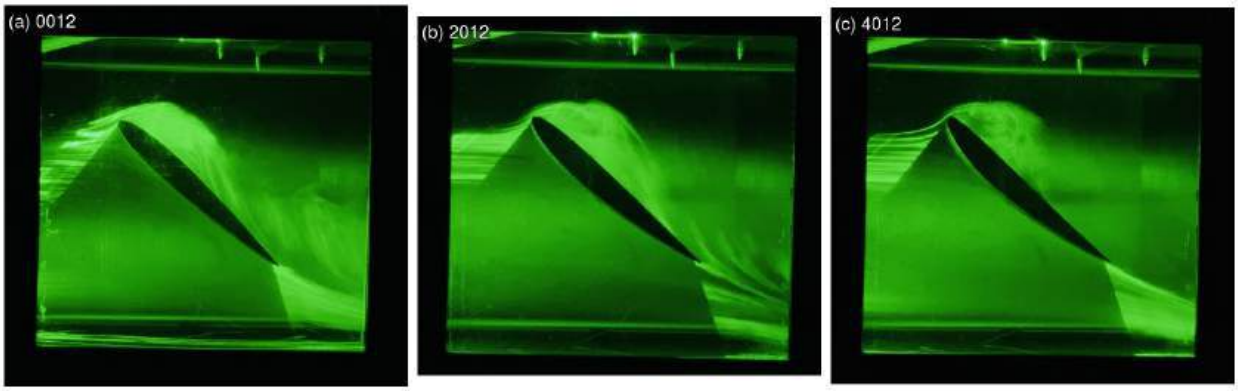


Fig 10: Deep stall at +31.4°: (a) NACA 0012, (b) NACA 2012, (c) NACA 4012. Large-scale separation covers the suction side in every case; the cambered sections show a coherent vortex rolling up from the leading edge.

4.3 Sensitivity to inlet conditioning

The NACA 0012 airfoil was re-investigated using the coarse grid replacing the fine grid (Fig. 11). The coarse-grid results show increased spread and less defined streaklines compared to the fine-grid experiment, which can be seen as the greater extent of the fog in (b). Given the flow velocities, the bar Reynolds numbers for the coarse and fine grids are approximately 140 and 70, respectively, both larger than the critical value (Re ≈ 50), above which grid bars generate unsteady wakes. These wakes decay on a characteristic length scale of 10-20 cell pitches, i.e., 14-28 cm for the coarse grid and 5-9 cm for the fine grid. Since the grid is placed 126.2 mm (4.968 in) upstream of the model, the wakes generated by the fine grid fully decay before interacting with the model in the test section, while the coarse-grid wakes still persist as turbulent structures visualized as the "fog" in panel (b).

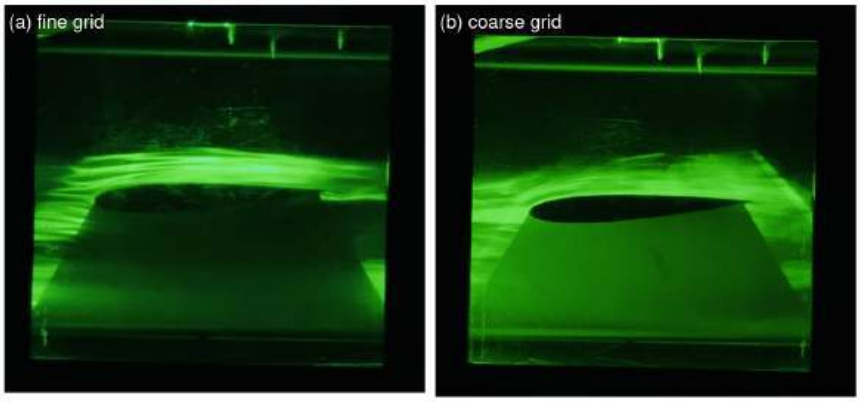


Fig 11: Sensitivity of the visualization to inlet conditioning: NACA 0012 with (a) the fine grid and (b) the coarse grid. Both photographs at the same angle and fan setting.

5. Discussion

The trends in Table 1 follow from circulation. Lift originates in the bound circulation the Kutta condition forces on the section, and for thin sections that mechanism fixes the lift-curve slope near 2π per radian regardless of the profile, which is why all five slopes cluster within a few percent of 0.110 per degree, and why thickness, which barely alters the circulation, barely alters the slope. Camber adds circulation at zero incidence rather than changing how circulation grows with angle, so it translates the lift curve without tilting it; the measured shifts of −2.0° and −4.1° reproduce the classical rule of roughly one degree per percent camber for this family (Abbott and von Doenhoff 124; Anderson 340). Thickness instead works through the pressure field: a fatter section demands stronger acceleration around the nose and a longer, steeper pressure recovery aft, which thickens the boundary layer and raises form drag. The doubling of zero-incidence drag across the thickness series, and the accompanying fall in peak L/D, is that mechanism rendered in numbers.

Against published measurement, the model's standing is mixed. At $Re = 3\times10^{6}$ the classical data give the 0012 a minimum drag coefficient near 0.006 and a maximum lift coefficient near 1.6 (Abbott and von Doenhoff); FoilSim at $Re = 1.3\times10^{4}$ reports 0.030 and 1.40. The drag figure is defensible: measurements below $Re = 10^{5}$ show section drag rising steeply as boundary layers stay laminar (Laitone 407; Selig et al. 22). The lift figure is not. At this Reynolds number, thin-airfoil stall arrives early; reported maximum lift coefficients for comparable sections sit near or below 1.0, with stall well under the model's 16° (Laitone 407; Selig et al. 22). The flow visualization agrees with the literature rather than the model: at +11.8° the suction surface of the 0012 already shows thickening and incipient trailing-edge separation (Fig. 7b), while FoilSim holds the lift curve linear for another four degrees.

Two internal failures surfaced during the sweeps, and both point at the model's architecture. The lift calculation is ideal-flow theory with corrections (NASA Glenn Research Center): it responds to geometry and angle, and, as the identical lift coefficients at $Re = 1.3\times10^{4}$ and 7.6×10^{5} showed, not at all to Reynolds number. The drag calculation is separate, built on stored empirical relations, and it is there that the symmetry violation lives: a symmetric section returned drag minima at −2° and up to 34% disagreement between mirrored angles, with the sign of the discrepancy not even consistent across the series, while the same runs kept lift antisymmetric to the last digit. A further geometric substitution compounds the drag comparison: the simulated shapes are Joukowski-family sections, not the NACA profiles they stand in for (Section 2.1). None of this is a failure in the program itself; it is simply a demonstrator that is designed to be operated in this way, as failures are hidden unless the user is willing to use negative angles of attack or change Reynolds numbers, both of which are not standard practice in a learning environment.

The practical conclusion for a student or preliminary designer is a division of labor. FoilSim ranks geometries correctly: thinner means less drag at zero lift, camber means lift without incidence, and the relative ordering of sections survives every check in this study. Its absolute numbers (drag levels, maximum lift, stall angle) should not be carried into any design at

low Reynolds number. A benchtop tunnel with a laser sheet and fog, at a total cost far below a single force balance, supplies the missing half: it cannot produce a coefficient, but it shows directly whether the flow state the model assumes actually exists on the geometry in hand.

6. Limitations

Computational. FoilSim Student Interactive Edition does not model any boundary layer, so all physics of separation, transition, and laminar-bubble interactions at $Re = 10^4$ are absent from it entirely; instead, its drag is defined by look-up tables, internal to the code, which are internally inconsistent; its sections approximate Joukowski profiles of the given nominal sections; its lift is independent of Reynolds number; and its span setting fixes the aspect ratio at 12.5, while the printed models have an aspect ratio of 0.46, so the finite-wing correction applied to the computed coefficients does not correspond to the wings actually tested. Coefficients are reported to two or three decimal places, and angles accepted as integers, which reduces the precision of fits to the lift-curve slope.

Facility. The test-section speed is established to within ±0.3 m/s (±20%), by visual tracking of streaks in 24 fps video footage vs. the fan speed rating; the Reynolds number inherits this uncertainty. The fine inlet grid does not fully dissipate the coherent structures, which appear as streaks in the videos, and the coarse grid is visibly inferior (Sect. 4.3). The view into the test section is further limited by scratches and reflections on the plastic office chair mat, from which the viewing window was cut. The fog seeding is not perfectly uniform, so the streakline density partly reflects the seeding density.

Models and mounting. The models occupy 28% of the test-section width; the flow around them is three-dimensional, and the mid-span illumination only partially suppresses the wingtip effects. The mounting post takes a visible fraction of the space beneath the model, redirecting the pressure-side boundary layer. The solid blockage increases with angle of attack to ~14% of the test-section area, at ±31.4°, reducing the freestream velocity over the model; the images from the deep stall thus show a confined flow, not free flight. The angles differ from the design ones by ±0.5°, due to the protractor precision. Only one side of the model is illuminated, and the runs are single, with the exception of the 0012 grid comparison. Thus, the repeatability of the setup is established for only one configuration.

Scope. The streaklines are not streamlines of the instantaneous flow in an unsteady domain: the flows depicted are all separated, and hence, unsteady, while the still images capture only one instant of it. No forces were measured; all quantitative statements are either computations or reference previous works, and the experiments only show which flow states occur.

7. Future work

The nearest improvements are measurement, not construction. A home-made anemometer or a 240 fps slow motion capture would reduce the Reynolds-number uncertainty severalfold with little effort. A simple force balance, even if it only uses a ivot arm on a digital

scale, would turn the qualitative comparison into a quantitative one, and multiple measurements at each condition would put error bars on the force coefficient. On the computational side, an XFOIL analysis of the real NACA coordinates at the same Reynolds number would isolate FoilSim's geometry approximation from its modeling approximations, and a pressure tap or smoke-wire flow visualization would localize the separation zone, not just reveal it. A more powerful fan would raise the Reynolds number, though reaching the literature's densest region would require about a hundred times the speed used.

8. Conclusion

The thickness and the camber were cleanly separated. The thickness had little effect on the lift curve but acted through the drag: for 9-15% thickness, it increased the zero-incidence drag by about a factor of two and decreased the peak value of lift-to-drag ratio from 14.0 to 12.3. The camber had no effect on the slope but caused a shift of approximately one degree per percent of camber on the zero-lift angle, and 4% of camber was enough to shift the section's maximum efficiency to zero incidence. The camber effect matched thin-airfoil theory quantitatively, while the thickness effect matched the literature, and the pictures also show that the cambered airfoil rotates the flow at zero angle, in the same direction as predicted by the calculation.

The two approaches justified the time spent on the analysis. The simulation gave numbers that a photo could not, while the photos captured the flow that the calculation could barely imagine: the flow separation beginning at +11.8°, deep stalls at ±31.4°, which the model did not reach, and the confinement and three-dimensionality of the real apparatus. Both methods cross-audited each other.

Moreover, the audit unearthed some interesting properties of the tools themselves: drag coefficients were not symmetric in the angle of attack for the symmetric profiles, lift coefficients failed to respond to a 58-times increase in Reynolds number, and the geometry of the model was only roughly similar to the one actually simulated. Thus, this ubiquitous educational tool can be trusted to reproduce the physics of circulation-induced lift and drag but should not be extrapolated to angles of attack beyond stall, to Reynolds numbers below $Re = 10^5$, or to cases where the geometry is not known to match the model. Defining these limits for the most accessible aerodynamics simulator is the goal of this study.

Works Cited

Abbott, Ira H., and Albert E. von Doenhoff. Theory of Wing Sections: Including a Summary of Airfoil Data. Dover Publications, 1959.

Anderson, John D. Fundamentals of Aerodynamics. 6th ed., McGraw-Hill Education, 2017.

JoDon3311. "DIY Open Return Wind Tunnel." Printables, www.printables.com/model/1500633-diy-open-return-wind-tunnel. Accessed Dec. 2025.

Laitone, E. V. "Wind Tunnel Tests of Wings at Reynolds Numbers below 70 000." Experiments in Fluids, vol. 23, no. 5, 1997, pp. 405–09.

NASA Glenn Research Center. "FoilSim Student Interactive Edition." NASA, www.grc.nasa.gov/WWW/K-12/FoilSim/. Accessed June 2026.

Selig, Michael S., et al. Summary of Low-Speed Airfoil Data. Vol. 1, SoarTech Publications, 1995.

Risk Factors For Skin Cancer By Navya Sampathkumar, Mentor By Dallin Lowder

Abstract

Skin cancer is a type of cancer that starts as a growth of cells on the skin. There are three types of skin cancer: melanoma, basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). It is the 5th most common cancer, affecting over a million people annually. This review will discuss the risk factors of skin cancer and how they affect the incidence of skin cancer globally. UV exposure is the main risk factor for all types of skin cancers. Reports show that people of color have a lower incidence of skin cancer than caucasians, but this may be due to underdiagnosis. Historically, males have a higher incidence of skin cancer than women. In the future, researchers should focus on inclusive, longitudinal studies of skin cancer to identify exactly how the risk factors affect its development, and doctors should be educated on how skin cancer appears on all types of skin conditions and complexions for better diagnosis.

Keywords

Biomedical and health sciences; genetics and molecular biology of disease; oncology; skin cancer; risk factors; melanoma; basal cell carcinoma; squamous cell carcinoma

Introduction

Skin cancer is a type of cancer that starts as a growth of cells on the skin. It is the 5th most common cancer, with an estimated 331,722 people diagnosed with melanoma globally. It can present itself in many ways, such as a sore on the skin or a growing mole. There are three types of skin cancer: melanoma, basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), with BCC and SCC being classed as keratinocyte skin cancer (KSC) since they originate from keratinocytes. Recently, it has been reported that there are rising incidence rates of melanoma and KSC globally, which have been attributed to the increased exposure to UV radiation. Moreover, evidence indicates that melanoma has had an increase of 3-7% in the White population over the past decade. Predominance of skin cancer observed in Oceania, North America and Europe. This might be linked to the fact that melanoma has proven to be closely linked to the color of one's skin and/or their geographical zone. In terms of epidemiology and gender, males have continued to have a higher risk of skin cancer than females, and the median tumour thickness of melanoma has also been higher in males. The risk of skin cancer can be raised by a variety of factors, but research concludes that prolonged exposure to UV radiation is the most prevalent risk factor (Leiter et al).

UV rays are a form of electromagnetic radiation that is emitted from the sun and other objects, such as tanning beds. They possess shorter wavelengths and higher energy than visible light, making them invisible to the human eye. The two types of UV that affect humans are UVA and UVB. UVA possesses longer wavelengths of 315 to 400nm compared to UVB; therefore, they are less harmful and only cause aging. UVB has short wavelengths from 200 to 315nm and plays a critical role in skin cancer. The sun produces 95% of UVA rays and 5% UVB rays, since

a lot is filtered out by the ozone layer (IARC, 35). While UV radiation can be beneficial for processes such as vitamin D production, it also plays a significant role in skin cancer. Figure 1 shows a simple illustration of how UV affects the skin barrier. This is because excess exposure to UV can cause mutations in the DNA which can cause skin cancer (P. Pfeifer, 3). Therefore, this type of radiation is categorized as a 'complete carcinogen' since it is a mutagen and a non-specific damaging agent (D'Orazio et al). With UV radiation having both properties of a tumor initiator and a tumour promoter, it can be linked to the three most common types of skin cancer: basal cell carcinoma, squamous cell carcinoma and malignant melanoma. The increased exposure to UV can be attributed to the rising incidence rates of skin cancer, as expressed above. Figure 2 shows the pattern of UV index forecast and the number of extremely high UV days and how that has changed over time, correlating positively with the incidence with skin cancer. The increasing exposure can be caused by a higher number of people visiting tanning beds, which use ultraviolet radiation to stimulate tanning. There seems to be a strong positive correlation between the use of tanning beds and the frequency of skin cancer. In fact, a significantly higher risk of melanoma was associated with 10 or more annual times of sunbed use. On the other hand, there are other risk factors that have also proved to have an impact on the risk of skin cancer.

Examples of other risk factors involved with increasing the risk of skin cancer include old age, race, gender, comorbidities, genetics and drugs. These risk factors play a role in the risk of skin cancer, but with varying degrees. For example, race is more likely to affect susceptibility compared to comorbidities. Furthermore, certain attitudes and behaviors have had positive correlations with risk of skin cancer (Dessinioti and Stratigos, 8886). For example, tanning and regularly spending time outside have been shown to play a role in increasing the risk of skin cancer. Different factors affect the risks of various types of skin cancer.

Identifying risk factors of skin cancer is extremely important because it helps in preventing, detecting and treating the disease early. Raising awareness about this issue will allow people to make better and informed choices about their lifestyle. It also improves the quality of research for treatments once the risk factors are thoroughly understood. Overall, educational awareness of the risk factors of skin has various positive effects.

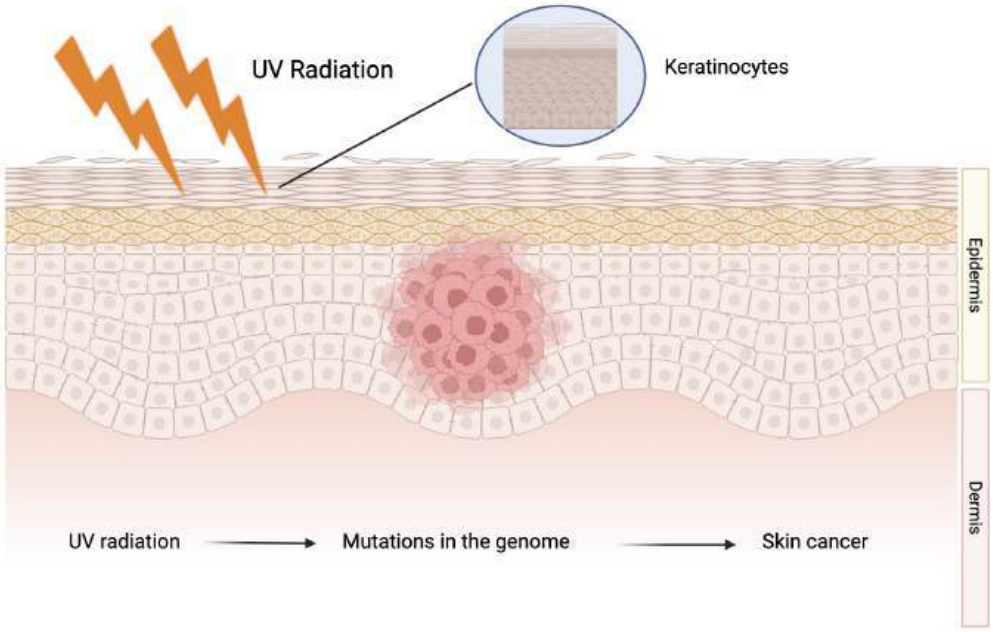
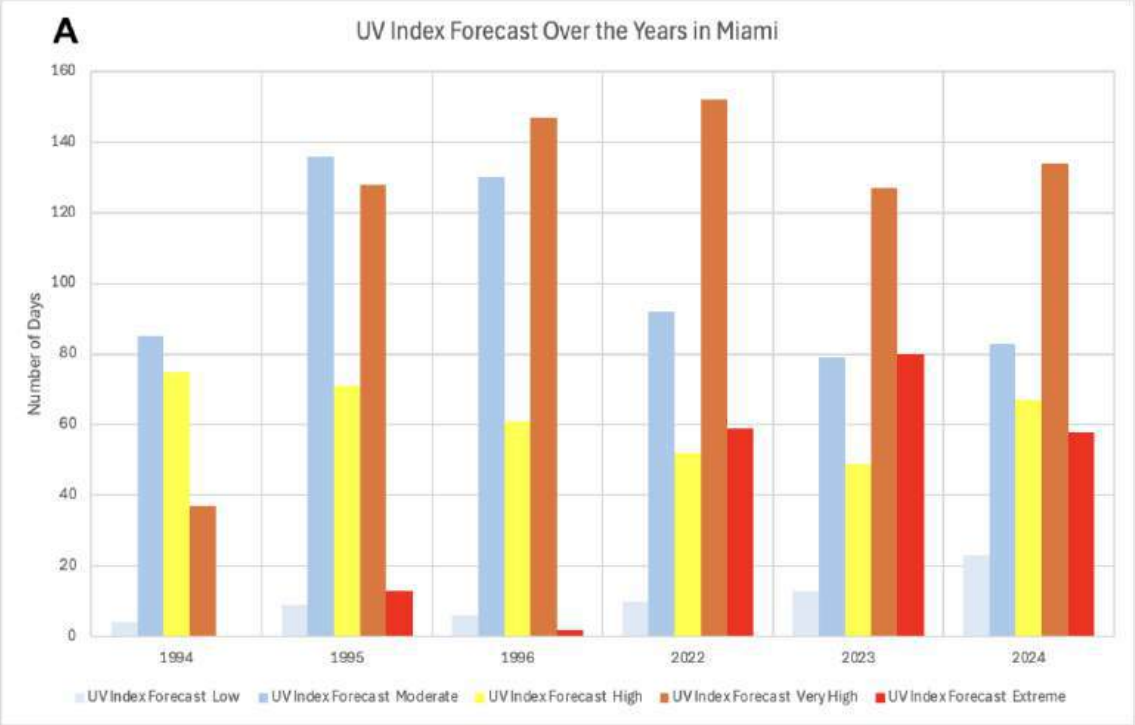
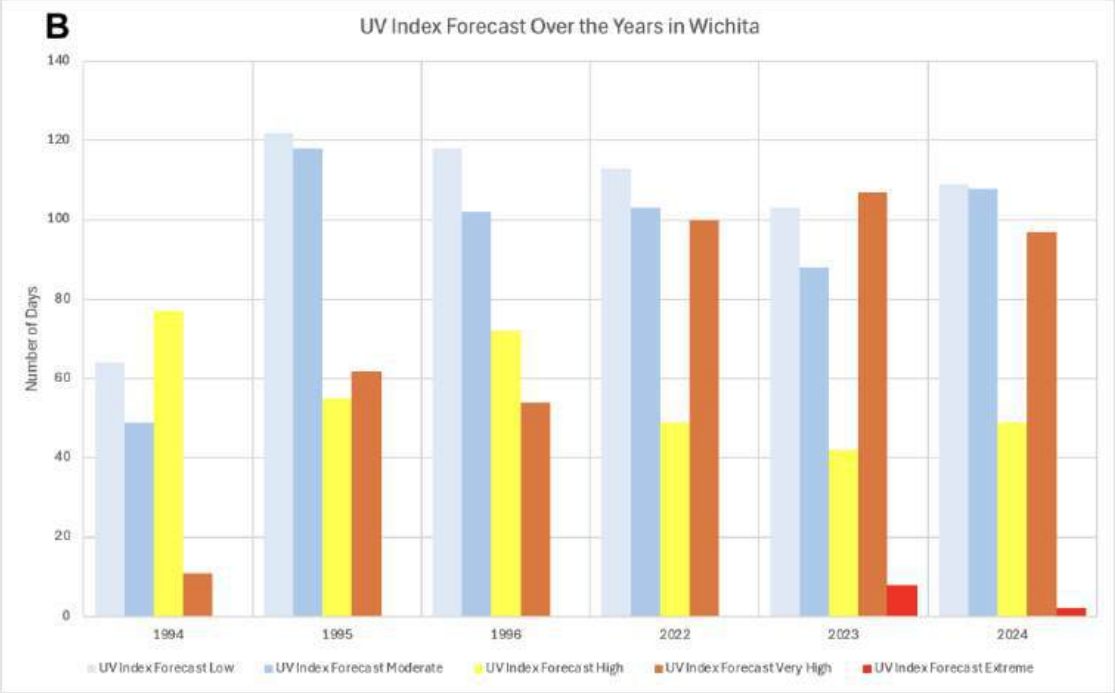


Figure 1: Shows the formation of skin cancer. UV radiation damages keratinocytes, which is the outermost layer of the skin barrier. This can cause mutations in the genome, which can lead to skin cancer.

UV's effect on different skin cancer types

UV leads to cell damage since it facilitates the formation of unwanted covalent bonds between adjacent bases in DNA, forming thymine dimers. However, our body repairs this damage with the DNA photolyase enzyme, which uses visible light to break the covalent bond holding the dimer together. However, in some cases, the repair can lead to mutations, which in turn result in skin cancer. Specific genetic conditions like xeroderma pigmentosum reduce the ability to repair DNA damage; therefore, individuals with this condition are at a significantly higher risk of developing skin cancer. Furthermore, each cancer type looks different on skin (Figure 3).


A
UV Index Forecast Over the Years in Miami
Number of Days
160
140
120
100
80
60
40
20
0
1994
1995
1996
2022
2023
2024
UV Index Forecast Low
UV Index Forecast Moderate
UV Index Forecast High
UV Index Forecast Very High
UV Index Forecast Extreme


B
UV Index Forecast Over the Years in Wichita
Number of Days
140
120
100
80
60
40
20
0
1994
1995
1996
2022
2023
2024
UV Index Forecast Low
UV Index Forecast Moderate
UV Index Forecast High
UV Index Forecast Very High
UV Index Forecast Extreme

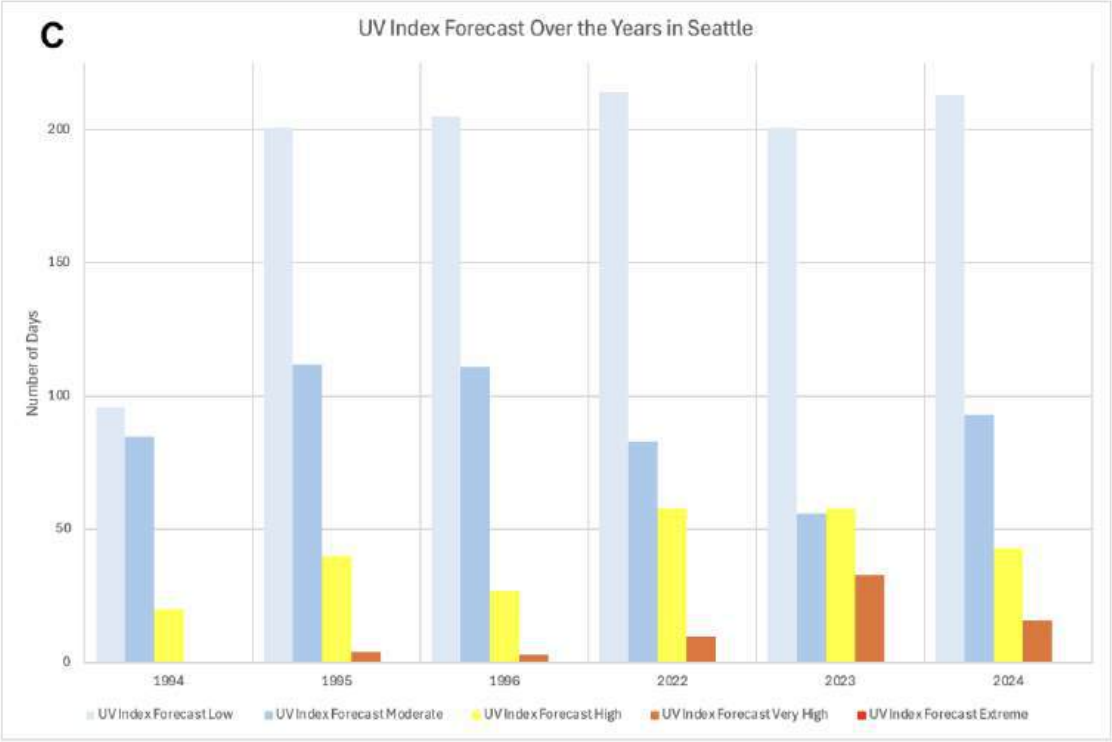


Figure 2: UV Index across time in three US cities. Shows the number of days of low, moderate, high, very high and extreme UV forecasts in the years 1994, 1995, 1996 and 2022, 2023, 2024 for the states of Miami, Wichita and Seattle. Graph A shows that the number of extreme UV days significantly increases from 1996 to 2022 and consistently remains that way for the following years. This also applies to graph B; however, the increase is not significant. Rather, here there are forecasted extreme UV days in 2023 and 2024, whereas in previous years, there were no extreme UV forecasted days. Graph C demonstrates no extreme UV forecasted across all years; however, the number of forecasted days for very high UV generally increases across the years.

Melanoma

Excessive exposure to UV leads to a higher risk of melanoma. In fact, the leading cause of melanoma is UV exposure compared to other factors such as race or gender. They can be caused by both short bursts of intense UV exposure and cumulative exposure over a longer period of time. That being said, attitudes and behaviors like using tanning beds can also increase the risk of melanoma since they add to the exposure to UV. Other factors, such as socioeconomic status and comorbidities, do not have as high a risk for skin cancer.

Basal cell

High UV exposure is the most prominent reason for BCC risk. Similar to SCC, this is also caused by a cumulative exposure to UV from childhood. Other factors have also been correlated with the risk of BCC, such as comorbidities like HPV, inactivation of specific genes like chromosome 9q22 and exposure to radiotherapy. Therefore, a variety of factors influence BCC; however, the major one is UV.

Squamous cell

Similar to melanoma, high UV exposure is the most prominent cause of developing SCC. The risk is cumulative, meaning that more exposure over time increases the chances of developing the cancer. Apart from UV exposure, there are other risk factors that significantly impact the risk of SCC, including genetic predispositions, immunosuppression and smoking. Another factor that is less prominent is age; however, it also has an impact on the risk of SCC.

Figure 3: The appearance of the 3 types of skin cancer on skin. The image on the left is melanoma, middle image is squamous cell carcinoma and right image is basal cell carcinoma. Open source images.

Race

Secondly, evidence has shown that race is a major factor in the risk of skin cancer, with most studies showing that non-Hispanic White individuals have a higher incidence rate of skin cancer compared to people of colour (Shao and Feng) . This is because White individuals have lower melanin levels, therefore are more susceptible to sunburns and skin damage. That being said, there is a significantly higher melanoma-specific death in non-Hispanic Black patients compared to their non-Hispanic White counterparts ($p<0.001$) (Brady et al). This mortality rate can be attributed to the late diagnosis and lack of awareness of skin cancer in people of color (Shao and Feng). In fact, the percent of melanoma cases initially diagnosed at Stage 3/4 was twice as high for Black and Native American patients compared to White patients. The reason for this delayed diagnosis can include clinical presentation, educational material, knowledge gap and the spread of misinformation such as the idea of darker skinned individuals not being susceptible to the risk of skin cancer. A study shows that 1,171/3467 non Hispanic White patients were diagnosed with cutaneous melanoma after 50 years of age (34%) whereas 19/73 non Hispanic Black patients were diagnosed after 50 years of age (26%). Although there were more non Hispanic White patients who were diagnosed, one cannot ignore the vast difference in sample size therefore, this study may not provide accurate findings (Shao and Feng). Still, incidence of the Asian, pacific island and Black population is increasing over the years likely to increase awareness, higher population and life expectancy increasing therefore more people get cancer which is why early diagnosis is extremely important (Gupta et al). This disparity suggests that public health frameworks focus too heavily on UV prevention, creating a wide diagnostic gap. Clinical training materials mostly showing skin cancer on lighter complexions leads to

healthcare providers being unequipped to recognize these malignancies when they deviate from the textbook in darker skin types. For example, for basal cell carcinoma, pigmentation is present in over 50% of the cases paired with an asymptomatic papule or nodule with translucent and central ulceration distributed in the head and neck. With squamous cell carcinoma, there is a scaling, well-circumscribed plaque or papule and ill-defined, rough, pink patches that may bleed or ulcerate. It is usually found in non-sun-exposed areas, including the lower body. It is also found in the arms, scalp and lip regions. Finally, with melanoma, there is an irregularly shaped, dark macule or patch from pigmented nevi. It often metastasizes early and then spreads to other parts of the body, and it is also found on non-sun-exposed areas such as the plantar, palmar, subungual skin and mucous membranes (S.Davis et al, 129). Ultimately, the lack of diagnosis seems to be a significant reason in addition to lower melanin levels.

Age

Incidence of basal cell carcinoma and cutaneous squamous cell carcinoma significantly increases above 65 years of age. The mortality rate is also higher because the treatment options are based on geriatric comorbidities and older individuals tend to get a reduced quality of treatment (Garcovich et al). This reduction in quality of treatment may result from bias and/or an emphasis in keeping the youth healthy rather than the aging population. A study was conducted with 552 participants aged between 70 and 93 years of age to find associations between skin cancer and its risk factors. Skin cancer was found in 25.5% of the participants and was more prevalent in males than females showing a statistically significant difference ($p<0.001$). Moreover, if patients had a previous cancer then they were 2.6 more times likely to develop skin cancer. Overall, specific factors were gender, outdoor work and socioeconomic status demonstrating that age is one of several interacting factors that can affect skin cancer risk (Sinikumpu).

Gender

Another significant factor of skin cancer includes gender. Generally, males continue to have a higher risk of skin cancer than females (Leiter et al). This can be attributed to factors like behaviors and genetics. In fact, variations in incidence rates between males and females are more likely to be attached to biological differences compared to behavioral attributes. For example, estrogen receptor beta is inversely associated with tumor aggressiveness while testosterone increases it (Schwartz, 112-118). Notably, external consumption of estrogen is correlated with an increased risk of skin cancer (Tang et al, 1469-1475). However, behavior also plays a role in the risk. A study surveyed 705 men recruited from Amazon Mechanical Turk and a majority of the men reported daily outdoor men but only few reported engaging in protective behaviors that reduced risk of skin cancer. 50.6% of men reported that their skin burns moderately when in the sun and 52.4% strongly agree that tanning makes them look more attractive. The study concluded that men consistently engaged in high risk behaviors for developing skin cancer (Adams et al). However, another study investigating alcohol consumption concluded that women

were at a higher risk of skin cancer than men when consuming alcohol (Wu et al). This presents a contradiction. This contradiction disappears when stratified by age. There is a distinct biological and behavioural shift, where women under 50 show higher rates of melanoma and BCC (a trend strongly correlated with cosmetic indoor tanning habits) whereas men over 60 dominate the statistics due to cumulative, lifetime sun exposure combined with age related changes in sex hormone production. This is seen in Table 1.

Cancer Type	Higher Risk (overall)	Higher risk (according to age)
Melanoma	Males	Women (<50 years)
Basal Cell Carcinoma	Males	Women (<40 years)
Squamous Cell Carcinoma	Males	Males (>60 years)

Table 1: While males face a higher overall lifetime risk across all three skin cancer types, epidemiological data shows a distinct change based on age. For example, women under the age of 50 experience higher rates of melanoma and basal cell carcinoma than their male peers

Comorbidities

Comorbidities is the simultaneous presence of 2 or more conditions present at the same time. This can have an impact on the risk of skin cancer. A study was carried out with 19,988 people and looked at the effects of socioeconomic status and comorbidities on freedom of recurrence in patients with treated non melanoma skin cancers. The results showed that individuals with diabetes had an overall of 3563 cases of skin cancer with the majority being BCC. Additionally, people with stroke or transient ischemic attack (TIA) reported a total of 928 cases. The study concludes that most individuals with comorbidities develop BCC in comparison to SCC apart from people with leukemia however the difference is low, with 69 cases being BCC and 70 cases being SCC (Ma et al). Therefore, comorbidities seem to have an effect on the type of skin cancer rather than the risk of developing it instead.

Genetics

Mutations in the genome can lead to skin cancer. Specifically, two genes that have shown a strong link with skin cancer include PTCH and p53. In fact, it has been observed that point mutations both in PTCH and p53 were predominantly UV specific transitions (Leiter et al). Genetic mutations within p53 have been shown to occur in nearly 50% of all cancers. They are especially common in skin cancers and are involved in the development of both NMSC and melanoma (Benjamin et al). This is problematic as p53 is the tumor suppressor protein and plays a vital role in preventing cancer therefore a mutation in this gene significantly increases risk of cancer. In the case of PTCH, mutations typically occur in SCC from individuals with a history of BCC (Ping et al). For melanoma, the PPP2R3B gene (on the X chromosome) is associated with

low melanoma outcome. This can also be linked to gender as well where females are reported to have lower rates of skin cancer than men and a reason for this can be because of PPP2R3B (Schwartz, 112-118).

Behavior

The lifestyle of a person can have an effect on their risk of developing skin cancer. Rather than working in isolation, these lifestyle habits directly interact with cellular genetics. For example, overconsumption of alcohol seems to be a risk factor. It has a dose dependent association with increased risk of BCC and SCC. This may be because alcohol directly decreases the protection offered by the skin through decrease in carotenoid concentration, hence reducing the skin's capacity to neutralize UV induced damage (Jain et al). When this behavioural vulnerability merges with an existing mutation in a tumor-suppressor gene such as p53, the cell's ability to repair DNA damage drops significantly. Food and other beverages also have an impact on skin cancer. A study was carried out in northern Italy with 380 cases of melanoma and 719 age-sex matched controls. There is a positive correlation between consumption of cereals, sweets and cabbages with melanoma risk. Conversely, there was a negative correlation between consumption of legumes, olive oil, eggs, onion and garlic and melanoma risk (Malagoli et al). Another factor that has an impact on risk of skin cancer is smoking. Interestingly, findings show that smoking can link to lower risks of melanoma and BCC. There has also been a weak inverse relationship between smoking and risk of developing all three skin cancers. An explanation for this can be nicotine's affinity for melanin developing tissues causing this inverse relationship (Dusingize et al). Finally, indoor tanning also links to risk of skin cancer as it uses UV radiation. Observations show that the higher the frequency of tanning beds, the higher the frequency of skin cancer. It is reported that there is a 1.8% increase in risk of melanoma for each additional session of indoor tanning per year and the development of multiple primary melanomas is 2.75x more likely with indoor tanning exposure (Dessinioti and Stratigios, 8894).

Conclusion

In conclusion, the risk factors for skin cancer include UV exposure, race, gender, behaviors, genetics and comorbidities. UV exposure is the main risk factor of all types of skin cancers. Reports show that people of color have a lower incidence of skin cancer than caucasians but this may be due to underdiagnosis. Historically, males have a higher incidence of skin cancer than women. Attitudes such as smoking, consumption of alcohol and tanning can affect the risk of developing skin cancer. Mutations in genes can increase the risk of skin cancer. Finally, the presence of comorbidities increases the risk of skin cancer.

Skin cancer is one of the most common types of cancers that affect over a million individuals annually. It is a painful disease that should be understood in order to find solutions for it. Understanding the risk factors of it allows for one to make more informed choices about their lifestyle and lower their risk of developing skin cancer. It also allows researchers to understand the disease better, and find better targeted treatments for it. Moving forward,

researchers should focus on inclusive, longitudinal studies of skin cancer to identify exactly how the risk factors affect its development. If this is pursued, it will help create more relevant solutions. Ultimately, looking beyond UV exposure requires a shift from generalized public health warnings to highly targeted clinical models. True progress in reducing skin cancer mortality depends on acknowledging the intersection of genetic risks, age demographics and diverse skin types into medical education. Additionally, doctors should be exposed and educated to how skin cancer appears on all types of skin conditions and complexions for better diagnosis.

Acknowledgments

I would like to thank my mentor, Dallin Lowder, for answering questions and providing accountability throughout my time writing this article.

Works Cited

Adams, Gabrielle J., et al. “Attitudes and Behaviors That Impact Skin Cancer Risk Among Men.” International Journal of Environmental Research and Public Health, vol. 18, no. 19, 23 Sept. 2021, p. 9989, 10.3390/ijerph18199989.

BENJAMIN, C, and H ANANTHASWAMY. “P53 and the Pathogenesis of Skin Cancer.” Toxicology and Applied Pharmacology, vol. 224, no. 3, 1 Nov. 2007, pp. 241–248, 10.1016/j.taap.2006.12.006.

Brady, Joshua, et al. “Racial Disparities in Patients with Melanoma: A Multivariate Survival Analysis.” Clinical, Cosmetic and Investigational Dermatology, vol. Volume 14, May 2021, pp. 547–550, 10.2147/ccid.s311694.

Davis, DiAnne S., et al. “Skin Cancer in Women of Color: Epidemiology, Pathogenesis and Clinical Manifestations.” International Journal of Women’S Dermatology, vol. 7, no. 2, Feb. 2021, 10.1016/j.ijwd.2021.01.017.

Dessinioti, Clio, and Alexander J. Stratigos. “An Epidemiological Update on Indoor Tanning and the Risk of Skin Cancers.” Current Oncology, vol. 29, no. 11, 17 Nov. 2022, pp. 8886–8903, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9689757/, 10.3390/curroncol29110699. Accessed 27 July 2026.

D’Orazio, John, et al. “UV Radiation and the Skin.” International Journal of Molecular Sciences, vol. 14, no. 6, 7 June 2013, pp. 12222–12248, https://www.mdpi.com/1422-0067/14/6/12222, 10.3390/ijms140612222. Accessed 27 July 2026.

Dusingize, Jean Claude, et al. “Genetic Variants for Smoking Behaviour and Risk of Skin Cancer.” Scientific Reports, vol. 13, no. 1, 6 Oct. 2023, p. 16873, https://www.nature.com/articles/s41598-023-44144-0#:~:text=Abstract, 10.1038/s41598-023-44144-0. Accessed 27 July 2026.

Garcovich, Simone, et al. “Skin Cancer Epidemics in the Elderly as an Emerging Issue in Geriatric Oncology.” Aging and Disease, vol. 8, no. 5, 2017, p. 643, 10.14336/ad.2017.0503.

Gupta, Alpana K, et al. “Skin Cancer Concerns in People of Color: Risk Factors and Prevention.” Asian Pacific Journal of Cancer Prevention : Apjcp, vol. 17, no. 12, 2016, pp. 5257–5264, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5454668/, 10.22034/APJCP.2016.17.12.5257. Accessed 27 July 2026.

International Agency for Research on Cancer. Solar and Ultraviolet Radiation. Www.Ncbi.Nlm.Nih.Gov, International Agency for Research on Cancer, 2012, https://www.ncbi.nlm.nih.gov/books/NBK304366/. Accessed 27 July 2026.

Jain, Neelesh P., et al. “The Effects of Alcohol and Illicit Drug Use on the Skin.” Clinics in Dermatology, vol. 39, no. 5, 21 May 2021, pp. 772–783, https://www.sciencedirect.com/science/article/abs/pii/s0738081x21000699?via%3dihub, 10.1016/j.clindermatol.2021.05.005. Accessed 27 July 2026.

Leiter, Ulrike, et al. “Epidemiology of Skin Cancer: Update 2019.” Advances in Experimental Medicine and Biology, vol. 1268, no. 1268, 2020, pp. 123–139, https://pubmed.ncbi.nlm.nih.gov/32918216/, 10.1007/978-3-030-46227-7_6. Accessed 27 July 2026.

Ma, Liqiao, et al. “The Impact of Socioeconomic Status and Comorbidities on Non-Melanoma Skin Cancer Recurrence After Image-Guided Superficial Radiation Therapy.” Cancers, vol. 16, no. 23, 1 Dec. 2024, p. 4037, https://www.mdpi.com/2072-6694/16/23/4037, 10.3390/cancers16234037. Accessed 27 July 2026.

Malagoli, Carlotta, et al. “Food and Beverage Consumption and Melanoma Risk: A Population-Based Case-Control Study in Northern Italy.” Nutrients, vol. 11, no. 9, 12 Sept. 2019, p. E2206, https://pubmed.ncbi.nlm.nih.gov/31547443/, 10.3390/nu11092206. Accessed 27 July 2026.

Pfeifer, Gerd P. “Mechanisms of UV-Induced Mutations and Skin Cancer.” Genome Instability & Disease, vol. 1, no. 3, 19 Mar. 2020, https://pmc.ncbi.nlm.nih.gov/articles/PMC8477449/, 10.1007/s42764-020-00009-8. Accessed 27 July 2026.

Ping, X. L., et al. “PTCH Mutations in Squamous Cell Carcinoma of the Skin.” The Journal of Investigative Dermatology, vol. 116, no. 4, 1 Apr. 2001, pp. 614–616, https://pubmed.ncbi.nlm.nih.gov/11286632/, 10.1046/j.1523-1747.2001.01301.x. Accessed 27 July 2026.

Schwartz, Matthew Robert, et al. “Sex Differences in Melanoma.” Current Epidemiology Reports, vol. 6, no. 2, 31 May 2019, pp. 112–118, https://www.healthaffairs.org/doi/pdf/10.1377/hlthaff.27.3.759, 10.1007/s40471-019-00192-7. Accessed 27 July 2026.

Shao, Kimberly, and Hao Feng. “Racial and Ethnic Healthcare Disparities in Skin Cancer in the United States: A Review of Existing Inequities, Contributing Factors, and Potential Solutions.” The Journal of Clinical and Aesthetic Dermatology, vol. 15, no. 7, July 2022, p. 16, https://pmc.ncbi.nlm.nih.gov/articles/PMC9345197/. Accessed 27 July 2026.

Sinikumpu, Suvi-Päivikki, et al. “Skin Cancers and Their Risk Factors in Older Persons: A Population-Based Study.” BMC Geriatrics, vol. 22, no. 1, 1 Apr. 2022, 10.1186/s12877-022-02964-1.

Tang, Jean Y, et al. “Menopausal Hormone Therapy and Risks of Melanoma and Nonmelanoma Skin Cancers: Women’s Health Initiative Randomized Trials.” JNCI: Journal of the National Cancer Institute, vol. 103, no. 19, 30 Aug. 2011, pp. 1469–1475, https://academic.oup.com/jnci/article-abstract/103/19/1469/900293?redirectedfrom=fulltext. Accessed 27 July 2026.

Wu, Shaowei, et al. “Alcohol Consumption and Risk of Cutaneous Basal Cell Carcinoma in Women and Men: 3 Prospective Cohort Studies.” The American Journal of Clinical Nutrition, vol. 102, no. 5, Nov. 2015, pp. 1158–66, https://pubmed.ncbi.nlm.nih.gov/26423390/, 10.3945/ajcn.115.115196. Accessed 27 July 2026.

Factors Explaining Differences Between Club and International Performances
By Abhinav Swaika

Introduction

Football players often compete in two environments : club and international football. Club football takes up most of a professional player's season, where they train and play regularly with the same teammates under consistent game tactics. While international football brings together players from different clubs, to play under one national team for a limited number of matches depending on the year. Because of these differences player performances vary depending on whether they are playing for club or country.

Analysts and fans often debate whether certain players perform better for their club or national teams. Some are fairly consistent scoring goals, creating chances, contributing to victories, at club level, but struggle to achieve the same in national teams vice versa. The difference in performance can be due to multiple factors such as team environment, tactical playstyles, team chemistry, time allocated to training and preparation with the team.

This study aims to examine whether there are measurable differences in football player performance between club and international football. By analysing statistical data such as goals,assists,pass accuracy, chance creation, then going further in depth with the use of regressive analysis, this research will compare the player performance in both environments and investigate possible reasons for any differences identified.

Research Question

Do professional football players' performances differ between club and international football?

Literature Review

The use of statistical analysis in football has increased significantly over the past two decades. Clubs, coaches and analysts increasingly rely on performance metrics to evaluate the current or past form. Such metrics that are used are goals and assists per 90 minutes, used to measure attacking contributions and provide a quantifiable method of comparing players in similar positions on performance across different competitions and teams.

Previous research and football analysts suggest that the preparation, team cohesion, tactical familiarity differ between club and national team football. According to The Analyst, club teams benefit largely due to longer time spent together by players, and larger number of games, than national teams. By being accumulated with the same environment, same tactical conditions, players are more accustomed to develop relationships and familiarity over time. In contrast, the national team meets only periodically throughout a season, which also depends on the international tournaments held during those years, resulting in some years offering very limited time to prepare with the team in order to play games. These structural differences may influence the players performance by affecting coordination, communication and more.

Physiological factors may also contribute to variations in performances. BelievePerform, goes in depth on the ongoing debate surrounding club and country football, suggesting that changes in environment, coaching relationships, and team dynamics affect players confidence and performance. Therefore, representing a national team can create immense pressure on players, due to increased public expectations and the symbolic meaning of the competition. At the same time some players receive greater motivation from these expectations, and perform beyond everyone's doubts leading to increased level of performances for their countries.

Despite the multiple differences that have been discussed, evaluating the performance between club and country football can still be challenging as team results alone cannot be enough statistical data to quantify the performances of individual football players. As it may not reflect the actual contributions of the players in the final result. Due to this many analysts and researchers rely on metrics such as goals and assists per 90 minutes to provide objective evidence of the players' improvements or fall in performances according to the team. Although previous research involving researchers and analysts has used data such as team cohesions, and psychological factors, little research has been conducted by directly comparing a players performance using metrics such as goals and assists per 90 minutes using large scale statistical data. Studies have been published on team performance, tournament outcomes rather than measuring the output players provide across two different environments.

This gap in the literature review allows for the foundation of this study. By analysing the statistical data of goals and assists of a large group of players around 100, from similar positions in both club and country games, this research will determine whether measurable differences exist between two competitive settings. Through quantitative analysis , this study will seek to contribute to a better understanding of how a change in environment, tactical differences and psychological factors affect a players performance.

Several limitations do propose themselves. International football has less matches played compared to club football, resulting in smaller sample sizes for some players. Additionally the limited games can prevent some players from showing their true output in those games, with the added differences in tactical objectives, environment, team quality. These factors cannot be controlled within this study but the large sample size is intended to reduce the differences across two environments to positively affect overall findings.

Methodology

This research uses a quantitative research design to investigate the relationship between player performances in club and international matches. Specifically it is to find the patterns of attacking output between club and country level, through regression analysis. By examining a large sample of players, the analysis will help in determining any patterns that appear across different competitive environments.

The data is taken from the publicly open source Fsbrief, which is a database on most professional football players, including their statistical data of goals and assists, in an arranged format allowing for easy comparisons between club and international football. The sample

consists of 100 players, in which 100 are solely forwards, from a variety of leagues and nationalities. For each player performance data was gathered separately for club and international football, and organized on an excel file. The variables include goals, assists, minutes played, matches played. To ensure that comparisons were more precise, the raw statistics were converted into 1000 per minute metrics.

The primary variables used are :

- Goals per 1000 minutes
- Assists per 1000 minutes

In which club performances are treated as the independent variable and international performances are treated as the dependent variable. By using 1000 minute metrics it will allow for a fair comparison, as the matches are limited in international football so subsequently minutes played would also be less in comparison.

Additionally, prior to the analysis all performance statistics were converted into 1000 minutes played metrics using the following equation -

Performance metric per 1000 = (Total metric / Minutes played) x 1000

This is done to compensate for the unequal playing time and enables more accurate comparisons between players.

Lastly players with insufficient time in either club and international football are excluded from the database to remove any possibilities of inaccuracy.

To improve the interpretability of the data of the results, all performance metrics were standardized at a rate of 1000 minutes played, done by the statistical software called STATA

Table 1

Variable	Obs	Mean	Std. dev.	Min	Max
club_go~1000	1,275	5.247977	3.885618	0	52.63158
club_as~1000	1,258	1.702122	1.774281	0	21.2766
national_g~0	659	5.533042	4.686255	0	31.25
national_a~0	651	2.015629	5.071204	0	100

The table above consists of all the variables included in this analysis. ‘Club_go~1000’ stands for the goals scored at club level in 1000 minutes on average. ‘Club_as~1000’, demonstrates the assists created in 1000 minutes at club level. The remaining variables represent equivalent statistics but for the international level of football played. It is important to note that Table 1, only demonstrates the descriptive statistics of all observations included in the dataset. Additionally the observations between club level and international data have a huge difference which is due to the unavailability of data for club players as they may not have represented their country as of yet in their career.

According to Table 1, the average goals scored per 1000 minutes (taken from the mean), is approximately 5 goals. The same goes with the performance at the international levels however the mean indicates a slight increase in assists moving from domestic to international leagues. This may indicate that moving to an international level, perhaps causes key forwards to play more of a goal scoring role rather than a creative player, which are results of different tactical approaches due to the limited number of games on the international level.

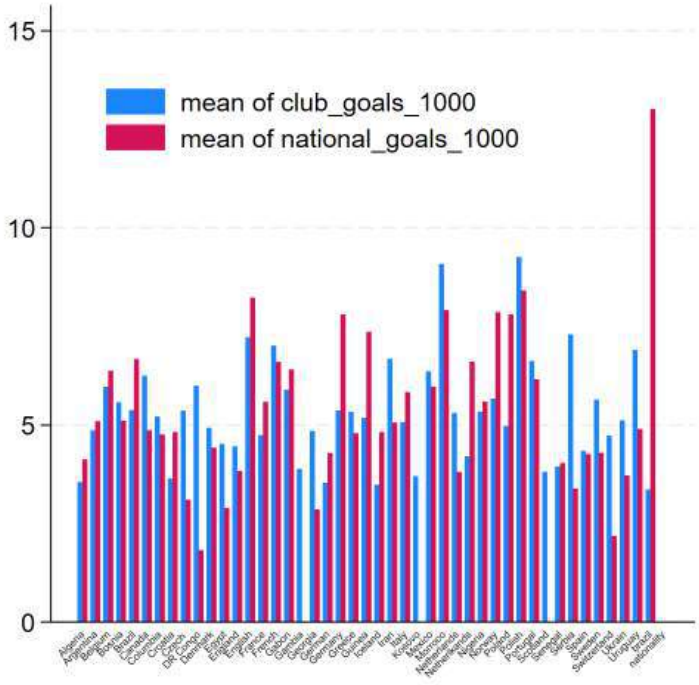


Figure 1: Mean goals per 1000 minutes at club and international by national teams

Figure 1 provides a visual comparison between the mean goals scored per 1000 minutes at club and international level for each national team. Table 1, summarizes the overall descriptive statistics, while Figure 1 highlights the variation across different national teams of mean goals, making patterns and trends easier to identify.

Table 2: t test (difference)

Variable	Observations (N)	Mean Difference	Std. dev.	t-value	Degrees Of freedom	p-value	Result	Confidence Interval (95%)
Goals (club-international)	614	0.6854846	5.024218	3.3808	613	0.0008	Significant	0.28 - 1.08
Assist (club-international)	614	-0.0307035	5.36195	-0.1407	603	0.8881	Not Significant	-0.45 - 0.39

Table 2 represents a t-test which is a statistical test that is used to determine if the difference between the two levels of competition is real or just occurred due to variation. This helps in identifying the differences in performance between club and international .

For goals, the results showed an average difference of 0.685 goals between club and international. With the standard deviation at 5.02, the t-value of 3.38 and 613 degrees of freedom. Additionally the resulting p-value is 0.0008 indicating that it is at a significant value. Since the difference is clearly not zero, the null hypothesis is rejected, proving that there is a significant difference in performance levels between club and national levels.

The confidence interval is at 95% and ranges from 0.28 to 1.08, indicating that there is a true average difference is likely positive, with players generally producing higher goal tallies at club level rather than international.

However, the data presented from assist conveys a different story. With the average difference being -0.0307, it indicates that approximately 0.03 assists are produced less at the club level compared to international games. And with the confidence interval ranging from -0.45 to 0.39, it is difficult to conclude whether the average difference is negative, positive or zero, therefore it cannot be confidently said that either environment produces more assist than the other. Additionally the p-value is above the significant value (which is 0.05), at 0.881, revealing that the null hypothesis is not rejected, and demonstrates that there is no statistically significant difference between club and international levels for assist.

Since the initial hypothesis has been proven wrong, as a forward player's overall performance does not increase which perhaps is due to a difference in tactical approaches or familiarity with the environment and people. To test this new hypothesis, a HHI index (Herfindahl-Hirschman index)was used, created by STATA, to test whether such factors are the cause of the differences revealed previously. In this case HHI is used as a proxy for concentration of national groups in a club, and is calculated by summing the squared proportions of players belonging to each nationality group within a squad. Higher HHI value would indicate greater nationality concentration while lower values will indicate teams more internationally diverse.

By using such a method, it will help in investigating whether players surrounded by a larger proportion of teammates with the same national backgrounds experience a positive result in performance, potentially due to benefits in communication, familiarity, and tactical understanding on the pitch.

Table 3: HHI Index OLS Regression

Outcome	Variable	Coefficient	Std. Error	p-value	R^2
Goals	Nationality HHI	0.386	0.399	0.333	0.035
Assists	Nationality HHI	-0.281	0.171	0.107	0.034

Table 3, demonstrates the nationality concentration within clubs, explaining the attacking differences. The club goal output is shown to have a positive relationship with concentration of

nationality as the coefficient is approximately 0.386, however this is not significant as the p-value is 0.333, much higher than the preferred range to be considered significant statistical data. Similarly the coefficient for assist is -0.102, suggesting a negative relationship, but as the p-value is large, it is also insignificant.

These findings overall suggest club nationality concentration does not significantly explain attacking performances.

Table 4: Fixed Effects Regression (Goals)

Variable	Coefficient	Std. dev.	p-value
Same-nationality teammate share	1.188	0.696	0.088
Age	1.285	1.278	0.309

Table 5: Fixed Effects Regression (Assist)

Variable	Coefficient	Std. dev.	p-value
Same-nationality teammate share	-0.908	0.403	0.027
Age	0.245	0.603	0.685

Table 4 & 5, have been constructed to examine whether a player's exposure to same nationality teammates influenced goal production. Unlike the HHI index measure, which captures overall squad nationality concentration, this variable measures the direct proportion of teammates who share the players nationality. The same was applied to the assist to determine any differences in creative output between club and international levels.

The coefficient is resulted as positive for goals at 1.188 suggesting that players with a greater proportion of same nationality teammates tended to generate a larger goal output. With this coefficient, it draws the conclusion that if the share of same countrymen goes up by 10% (0.1) then the club goals would go up by (0.1*1.18 = 0.19) or ≅ 2 goals per 1000 minutes , which is a significant increase. Additionally the p-value of 0.088 making the statistical data significant. From the information in table 5, the coefficient is -0.908 suggesting that players with a higher proportion of same nationality teammates produce lesser amount of assist output, therefore it proves that same nationality teammates do not improve creative output, but do result in a significant increase in goal output

Conclusion

Overall the results provide strong evidence that international management teams appear to operate under different tactical and structural conditions compared to clubs . The data suggests that there is a significant increase in goal output at club level compared to international games.

Multiple factors can contribute to this, one of them was the HHI index which proved that a higher proportion of same nationality players tend to create a higher goal output almost by 2 goals per 1000 minutes. Such an increase is perhaps due to the role required for attacking players. At the international level clubs are required to keep an acceptable standing, and win at a constant rate, additionally the club football provides significant opportunities giving players massive amounts of time to commit themselves both emotionally and physically to all the facilities and people around them. With such time spent, a trust is built and so forward players can focus on what best suits them to positively contribute to the team, which is to score goals. Conversely, at the international level there is a massive dip in time spent familiarizing with the team, and are only given a limited amount of time to display themselves to the best ability in order to win favour, or impress the management team. Due to such factors forwards are perhaps required to play more of a playmaking role, and are instructed to play more of a team, track back to defend, in order to get everyone in the team to contribute both offensively and defensively. Additionally forwards, perhaps don't have enough chemistry with teammates, as the regression analysis proved that even with same nationality teammates, the goal output decreases, but the assist output increases due to change in requirements. This suggests that because of the change in tactical instruction, forwards have more opportunities on the ball allowing for more chances to be created by them.

To fix this issue of lack of goals at the international level, international management teams can increase team building exercises at national camps and practice sessions. But the main factor that contributes to this problem is the lack of time and games during the season, which cannot be increased due to tight scheduling around a club season. But perhaps management teams for national teams can start including players who have abilities to adapt extremely quickly and contribute in multiple areas of play, to try to convert all the skills of the chosen players into a positive momentum, giving them a boost in performance compared to opposition international teams.

To conclude, this research found that attackers goal output increased compared to international levels while the assist output remained fairly similar but slightly in the international levels favour and included the variation across national teams. These findings suggest that while individual attacking performance is relatively consistent across competitive environments, contextual factors influence creative output across professional players. Further research into how tactical systems, teammate quality, playing style affect the relationship of performance to the two levels of competitive football may be important to figure out a proper solution. Additionally expanding the analysis to midfielders and defenders using different performance metrics will help in creating a more comprehensive understanding of player performance.

Work Cited

Forman, Sean. “Football Statistics and History | Fbref.Com.” Sports Reference, Oct. 2004, fbref.com/en/.

Nordvik, Christer. “Fotmob - Football Live Scores.” Fotmob, 2004, www.fotmob.com/.

Whitemore, Jonny. “5 Key Differences between International and Club Football | Opta Analyst.” Opta Analyst, 21 July 2021, theanalyst.com/articles/five-differences-between-international-and-club-football.

“Best Strikers (ST) in FC 26 | FC Ratings.” FC Ratings, 11 Sept. 2023, www.fcratings.com/positions/striker-25.

Tylkin, Igor. “Prove Them Wrong: Do Professional Athletes Perform Better When Facing Their Former Clubs? - Sciencedirect.” Science Direct, June 2022, www.sciencedirect.com/science/article/abs/pii/S2214804322000532.

Becklas, Andre. “FIFA World Cup | Kaggle.” Kaggle, 2018, www.kaggle.com/datasets/abecklas/fifa-world-cup.

Passingham, Alexa. “Club vs. Country: The Psychological Debate - Believeperform - the UK’s Leading Sports Psychology Website.” Believe Perform, members.believeperform.com/club-vs-country-the-psychological-debate/. Accessed 27 June 2026.

Financial Sustainability, Regulation, and Competitive Balance in European Football Leagues By Hashem El Barbir

Abstract

European football has undergone a dramatic financial transformation over the past two decades, fuelled by growing broadcasting deals, commercial partnerships, and record-breaking transfer fees. While the biggest clubs now earn revenues that were unimaginable a generation ago, this raises an important question: does financial growth come at the expense of fair competition? This paper investigates how financial growth has affected club sustainability and competitive balance in the Premier League, arguing that elite clubs have benefited the most while UEFA's Financial Fair Play (FFP) regulations have further widened the gap between rich and poor clubs. To explore this, secondary data on club revenues, debt, transfer spending, and league points were analysed using k-means clustering to compare financial and performance groupings across pre- and post-FFP periods. The results partially supported the argument that the FFP has actually widened the gap between rich and poor clubs. Clubs grouped closely along financial lines, with higher-revenue clubs generally finishing higher in the league, though some mid-revenue clubs still competed at a similar level to wealthier rivals. While financial growth and regulation have improved overall stability, they also appear to be locking in long-term inequality, raising questions about whether FFP is actually achieving its goal of a more competitive game.

Keywords UEFA Financial Fair Play, Premier League, club financial sustainability, competitive balance, revenue inequality, transfer expenditure, wage spending, k-means clustering.

Introduction

Modern European football is increasingly shaped by financial power as much as by sporting performance. Record-breaking revenues, rising transfer fees, and growing wage bills have transformed football clubs into global entertainment businesses, raising concerns about financial sustainability and competitive balance across leagues. In response to these developments, regulatory frameworks such as UEFA's Financial Fair Play (FFP) (UEFA, 2012) were introduced to limit excessive losses and promote long-term financial stability. UEFA's Financial Fair Play (FFP) is a regulatory framework introduced in 2011 that requires clubs competing in European competitions to demonstrate that their spending does not significantly exceed their income over a rolling three-year period. Despite these interventions, large financial disparities between clubs persist, prompting debate over whether increased investment strengthens competition or entrenches the dominance of elite teams.

This paper investigates the following research question: How does financial investment and UEFA's Financial Fair Play regulations affect club sustainability and competitive balance in the Premier League? It is hypothesised that increasing financial investment has improved financial performance and sustainability for elite clubs, while simultaneously reducing

competitive balance by reinforcing the dominance of high-revenue teams over lower-income clubs. We find that FFP has restricted the ability to invest in players and infrastructure, therefore increasing the imbalance, but different groups of teams have been affected in different ways.

A key finding across the literature is that wage spending is one of the strongest predictors of league position, meaning financial resources translate directly into competitive outcomes (Szymanski & Kuypers, 1999). This relationship has been reinforced by the growing concentration of broadcasting revenues among top clubs, which can be measured through the Gini coefficient, a statistic used to show income or revenue inequality within a group. A higher Gini coefficient indicates that revenue is becoming more concentrated among fewer clubs, suggesting that broadcasting income is not being distributed evenly across the league. This has made it structurally harder for smaller clubs to compete regardless of how efficiently they operate (Pawlowski et al., 2018). Deloitte's annual Football Money League reports have consistently shown that the same group of clubs dominate revenue rankings year after year, suggesting that financial hierarchies in football are becoming increasingly entrenched (Deloitte, 2024). Kuper and Szymanski (2009) further demonstrated that clubs which consistently outspend their rivals tend to dominate league standings over the long run, while Franck (2014) argued that without regulation, football risked a financial doping dynamic where clubs took on unsustainable losses simply to buy short-term success, pointing to the collapses of Leeds United and Portsmouth as real consequences of unchecked spending. On the regulatory side, while FFP was introduced to prevent exactly this kind of behaviour, several studies suggest it may have done the opposite, protecting the advantages of already-wealthy clubs while restricting others from investing and catching up (Müller et al., 2017). Peeters and Szymanski (2014) reinforced this concern, showing through economic modelling that FFP's break-even requirement disproportionately benefits clubs already earning high revenues, since they can sustain higher spending within the rules. What matters here is whether financial regulation is actually making football more competitive, or simply making it harder for smaller clubs to ever challenge the teams already at the top.

While existing research has examined financial inequality and FFP broadly, fewer studies have directly compared clubs based on their financial characteristics over a defined pre- and post-FFP period. This paper contributes by clustering ten Premier League clubs between 2010 and 2024 according to financial and performance data. The findings suggest that elite clubs have consolidated their dominance under FFP, while mid-table and lower clubs have faced increasing financial pressure, adding to the growing body of evidence questioning whether FFP achieves its predicted goals.

Section 2 outlines the methodology and data sources. Section 3 presents the results. Section 4 discusses the findings in relation to the hypothesis and literature. Section 5 concludes with key findings, limitations, and directions for future research.

Methodology

This study uses secondary financial and performance data from ten Premier League clubs across three competitive tiers (top, mid-table, and lower) comparing figures from 2010 and 2024. These years were selected because 2010 represents the period just before the introduction of UEFA Financial Fair Play regulations, while 2024 reflects a mature stage where FFP has been fully implemented, allowing for a long-term comparison of its impact. The variables collected include club revenue, wage expenditure, and league points. These variables were selected because they directly capture both the financial input a club has available and the competitive output it produces, making them the most relevant measures for examining the relationship between financial investment and on-pitch performance. These were sourced from publicly available reports including the Deloitte Football Money League (https://www.deloitte.com), Swiss Ramble (https://swissramble.blogspot.com), Transfermarkt (https://www.transfermarkt.co.uk), and club annual accounts filed at Companies House (https://www.gov.uk/get-information-about-a-company).

To analyse the data, k-means clustering was applied, a widely used statistical method for partitioning observations into groups based on similarity (MacQueen, 1967). A value of k=3 was selected to reflect the three competitive tiers of clubs (top, mid-table, and lower), aligning the statistical grouping with real-world league structure while maintaining simplicity. This number was also chosen because clustering by financial characteristics alongside performance allows the analysis to infer more than just on-pitch outcomes, it captures underlying financial patterns as well. Different values of k were tested, but three clusters proved sufficient to capture meaningful differences between groups without overcrowding the analysis, where too many clusters would make patterns harder to interpret.

Clustering financial variables (revenue, wages, net spend, debt) independently from performance variables (revenue, league points) allows for a direct comparison between the two, specifically, whether clubs that cluster together financially also cluster together in terms of on-pitch outcomes. Revenue is included in both groupings because it acts as a bridge between the two — while it is fundamentally a financial measure, it is also heavily influenced by performance, since better league finishes and European qualification directly increase a club's prize money and broadcasting income. Cluster membership between the two different seasons can give information about the effect of FFP on clubs. This approach was chosen because it identifies structural groupings, which is more useful for examining how FFP has affected different types of clubs. All data was standardised before clustering to ensure no single variable dominated the results.

Results

The data collected for this study shows a clear increase in financial activity across all ten clubs between 2010 and 2024, but this growth was also unevenly distributed. Average revenue across the sample rose from £151m in 2010 to £363m in 2024, a 140% increase over the period. Average wages grew from £97m to £267m, and average debt rose from £161m to £334m. While part of this growth reflects general inflation over the 14-year period, a 140% rise in revenue still

significantly outpaces the roughly 55% cumulative inflation seen in the UK over the same time, suggesting the increase reflects genuine financial expansion rather than just the changing value of money.

This growth, however, was far from evenly distributed. Manchester City's revenue grew by £546m over the period, while Burnley's grew by just £70m. This is not just a numerical difference, it reflects a growing gap between clubs that have the commercial power to benefit from football's financial expansion, and those that do not. The Premier League has grown as a product, but the clubs within it have not grown at the same rate.

Looking at revenue growth in proportional rather than absolute terms tells a slightly different story. Arsenal's revenue grew by £297m in absolute terms, but this represents only a 116% increase relative to their 2010 base. Brentford, by contrast, grew by just £162m in absolute terms, yet this amounts to a massive 1,800% increase relative to where they started, largely because their promotion to the Premier League gave them access to much larger broadcasting, commercial, and matchday revenues.This is significant because FFP's break-even rule is fundamentally proportional, clubs are allowed to spend in line with what they earn, not a fixed amount. This means a club like Brentford, despite growing rapidly in percentage terms, is still working with a much smaller absolute spending capacity than Arsenal, who only needed modest proportional growth to maintain a far larger financial advantage. In this sense, FFP does not punish proportional growth, but it also does nothing to close the absolute gap between clubs that were already large in 2010 and those that were not.

Revenue: 2010 vs 2024 (£m)

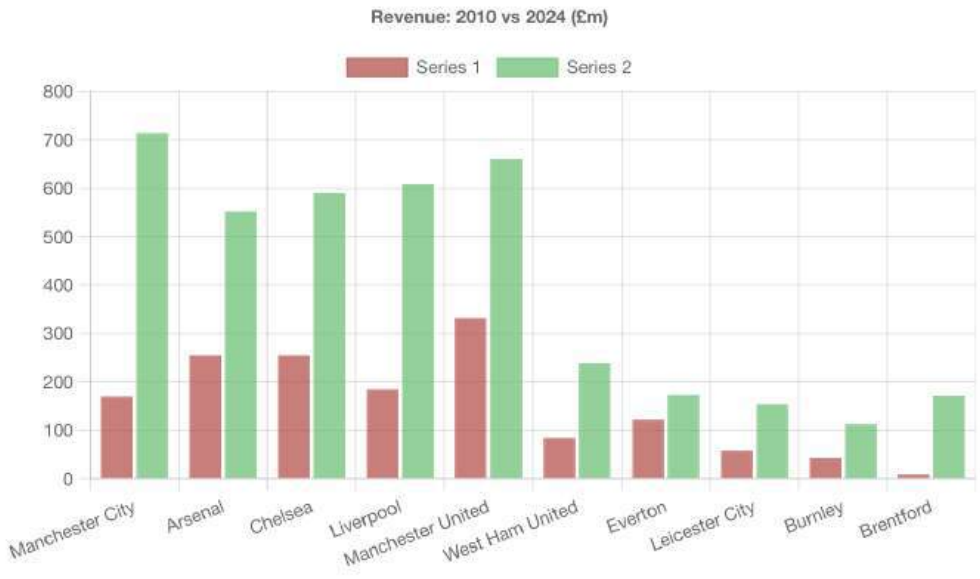


Table 1. Premier League Club Financial and Performance Data: 2010 and 2024 Compared

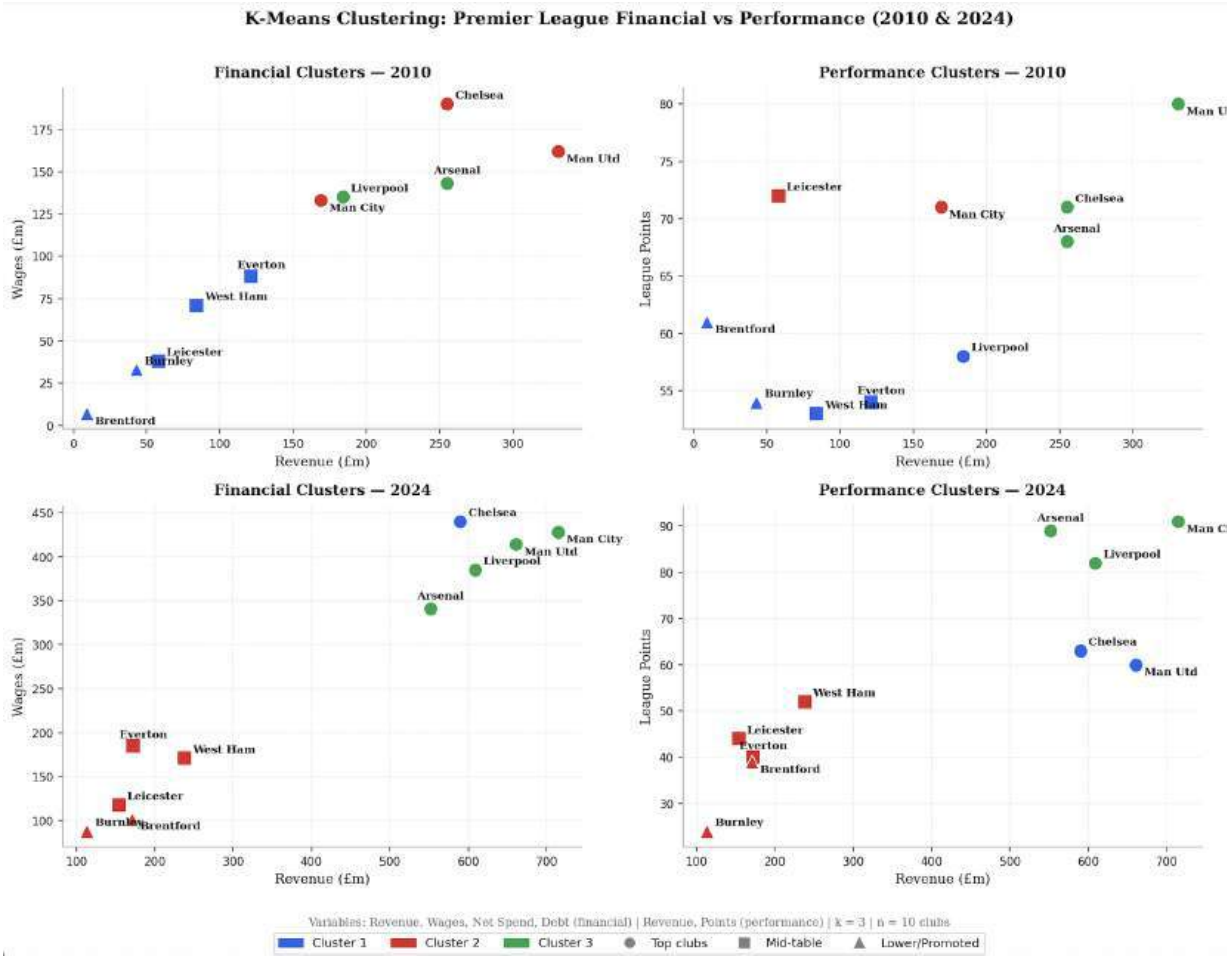
Selected clubs across three competitive tiers. All financial figures in £ millions.

		2010 / 2010-11 Season					*2024 / 2023-24 Season*					*Change (2010→2024)*	
Club	**Tier**	**Revenue**	**Wages**	**Net Spend**	**P&L**	**Debt**	**Revenue**	**Wages**	**Net Spend**	**P&L**	**Debt**	**Δ Revenue**	**Δ Points**
Manchester City	*Top*	169	133	+80	(197)	352	715	428	+55	+27	302	**+546**	**+20**
Arsenal	*Top*	255	143	(38)	+15	264	552	341	+144	+35	229	**+297**	**+21**
Liverpool	*Top*	184	135	+14	(49)	89	609	385	+75	+28	259	**+425**	**+24**
Chelsea	*Top*	255	190	+57	(68)	764	590	440	+337	(178)	817	**+335**	**−8**
Manchester United	*Top*	331	162	+51	+12	308	661	414	+75	(113)	622	**+330**	**−20**
West Ham United	*Mid*	84	71	+4	(11)	27	238	171	+49	(23)	214	**+154**	**−1**
Everton	*Mid*	121	88	+18	(8)	45	172	185	+49	(68)	497	**+51**	**−14**
Leicester City	*Mid*	58	38	+4	(2)	12	154	118	+6	(31)	195	**+96**	**−28**
Burnley	*Lower*	43	33	+2	(1)	6	113	88	+88	(81)	142	**+70**	**−30**
Brentford	*Lower*	9	7	+1	0	2	171	102	+15	(16)	68	**+162**	**−22**

Note: P&L = pre-tax profit/loss. Net Spend = transfer expenditure minus transfer income. Losses and negative values shown in brackets. Brentford's 2010 figures reflect Championship-level operations prior to Premier League promotion.

Looking at the clustering analysis, three distinct groups emerged across both financial and performance dimensions in 2010. Financially, the low cluster contained Burnley and Brentford, the medium cluster included West Ham, Everton, Leicester, Liverpool, and Man City, and the high cluster contained Arsenal, Chelsea, and Manchester United, with average revenues of roughly £26m, £151m, and £296m respectively. It is worth noting how close together these clusters were compared to 2024. The performance clusters in 2010 also did not fully match the financial ones. Brentford and Leicester both appeared in a higher performance cluster relative to their financial position, suggesting that in 2010 clubs could still compete above their financial level to some degree. This is an important starting point, because it shows that the link between money and results, while present, was not yet as fixed as it would later become.

Looking at debt as a proportion of revenue also reveals an important shift. In 2010, Chelsea carried the highest debt-to-revenue ratio in the sample at roughly 300%, while most other clubs sat well below 100%. By 2024, Manchester United's debt had reached 94% of their revenue and Everton's had climbed to nearly 290%, despite Everton earning significantly less than the top clubs. What this suggests is that financial pressure has not stayed at the top, it has spread downward, with mid-table clubs now carrying debt burdens relative to their income that are comparable to what only the biggest clubs were carrying in 2010.



By 2024, the financial picture had changed quite a lot. The high cluster now included Man City, Arsenal, Liverpool, Chelsea, and Manchester United, averaging £625m in revenue, while the low cluster of Leicester and Burnley averaged around £134m. The gap between the top and bottom clusters had grown from roughly £270m to £491m, and the overall spread of clubs had widened significantly. This is relevant to the research question because it suggests that FFP, rather than reducing financial differences as it was intended to do, has coincided with a period where those differences have actually increased. Clubs that were already earning strong revenues when FFP was introduced in 2010 were in the best position to grow within the rules, since FFP essentially allows clubs to spend in proportion to what they already earn. For clubs lower down, this created a ceiling where catching up through heavy investment was no longer really an option without risking punishment.

This is perhaps most clearly shown through the debt figures. Chelsea's debt rose from £764m to £817m and Manchester United's from £308m to £622m, despite both operating under FFP throughout the period. This suggests FFP has not removed financial risk at the top, it has just shifted what form it takes, away from direct operating losses and towards ownership financing and infrastructure costs which fall outside the regulations. Everton is a useful contrast here. Their debt grew from £45m to £497m as they tried to grow the club and build a new

stadium, and they ended up receiving a points deduction for FFP breaches in 2023. The difference in how financial risk has played out for a top club versus a mid-table one is one of the more telling patterns in this dataset.

The performance clusters in 2024 aligned much more closely with the financial clusters than they had in 2010, which is meaningful because the k=3 model was chosen to reflect the study's three broad club categories: elite, mid-table, and lower-revenue clubs. This choice captured more detail than a simple two-cluster model while avoiding the over-crowding of four clusters, and the findings suggested that the 2024 groups were more clearly separated. In 2010 there was a clear gap between where clubs stood financially and how they actually performed on the pitch. By 2024 that gap had largely closed, with financial position becoming a much stronger indicator of league outcomes. The main exception was Chelsea, who despite the highest net spend of any club at £337m finished sixth with 63 points. This shows that spending without a clear strategy does not automatically produce results. Arsenal by contrast grew their revenue from £255m to £552m, kept their finances relatively controlled, and finished second with 89 points, which is arguably the most efficient financial-to-performance conversion in the sample. These differences suggest that while the overall trend points toward financial power shaping competitive outcomes, how clubs manage their spending within that still matters. The hypothesis is therefore partially supported, the growth in investment has strengthened the position of elite clubs and FFP has made it harder for others to challenge them, but how clubs spend is still just as important as how much.

Conclusion

This paper set out to examine how financial investment and UEFA's Financial Fair Play regulations have affected club sustainability and competitive balance in the Premier League between 2010 and 2024. Using k-means clustering across financial and performance variables for ten clubs, the analysis found that financial inequality has widened significantly over this period, and that financial position has become a much stronger predictor of league outcomes than it was before FFP came into effect. The gap between the top and bottom financial clusters grew from roughly £270m in 2010 to £491m in 2024, and the alignment between financial and performance clusters tightened considerably, suggesting that the link between money and results has become more fixed over time. This provides partial support for the hypothesis, financial investment has clearly reinforced the dominance of elite clubs, and FFP appears to have made it harder for lower-revenue clubs to close the gap, but the Chelsea and Arsenal contrast shows that how clubs spend still matters alongside how much they spend.

A more unexpected finding was how financial risk has spread downward. In 2010, debt burdens relative to revenue were concentrated at the top of the league, but by 2024 mid-table clubs like Everton were carrying debt-to-revenue ratios comparable to what only the biggest clubs carried a decade earlier. This points to a deeper structural issue. FFP has not removed financial risk from football, it has shifted where that risk sits and what form it takes, while leaving smaller clubs more exposed when they attempt to grow.

There are limitations to this study. The sample of ten clubs across two seasons captures broad structural patterns but cannot fully reflect year-by-year volatility, ownership changes, or the impact of one-off events such as the COVID-19 period on club revenues. The analysis also relies on secondary financial data which, while publicly available, can vary in how it is reported across sources. Future research could expand the sample to include more clubs across multiple European leagues, track financial and performance variables on an annual rather than two-point basis, and incorporate qualitative factors such as ownership models and recruitment strategy, which this paper was not able to capture quantitatively.

Overall, the findings raise genuine questions about whether FFP is achieving its stated goal of promoting financial fairness and competitive balance. The regulation has improved overall stability and reduced the risk of clubs collapsing in the way Leeds and Portsmouth once did, which is a meaningful achievement. But it has also coincided with a period where the financial and competitive gap between the top and the rest of the league has grown, not shrunk. If the goal of FFP is to make football more competitive, rather than simply more stable, the evidence from this paper suggests it is not currently succeeding.

Works Cited

Birkhäuser, Stephan, et al. "Did UEFA's Financial Fair Play Harm Competition in European Football Leagues?" Review of Managerial Science, vol. 13, no. 1, July 2017, pp. 113–45, https://doi.org/10.1007/s11846-017-0246-z.

"Burnley Finances 2023/24 — Swiss Ramble." Substack.Com, 2025, https://swissramble.substack.com/p/burnley-finances-202324.

Carmichael, Fiona, et al. "Maintaining Market Position: TEAM Performance, Revenue and WAGE Expenditure in the English Premier League." Bulletin of Economic Research, vol. 63, no. 4, Aug. 2011, pp. 464–97, https://doi.org/10.1111/j.1467-8586.2009.00340.x.

City, Man. MCFC Annual Report 2010-11. Man City, 18 Nov. 2011, https://www.mancity.com/news/club-news/club-news/archive/2011/november/mcfc-annual-report.

"Comparative Analysis of Competitive Balance in the English and Spanish Top-Flight Leagues ." Journal of Physical Education and Sport, vol. 25, no. 2, University of Pitești, Feb. 2025, https://doi.org/10.7752/jpes.2025.02039.

Cox, Adam, and Christina Philippou. "Measuring the Resilience of English Premier League Clubs to Economic Recessions." Soccer & Society, vol. 23, no. 4–5, Apr. 2022, pp. 1–18, https://doi.org/10.1080/14660970.2022.2059858.

Deloitte. "Deloitte US | Audit, Consulting, Advisory, and Tax Services." Deloitte United States, 2023, https://www.deloitte.com.

"Everton: Club Deducted 10 Points for Premier League Profit and Sustainability Rule Breach." Sky Sports, https://www.skysports.com/football/news/11095/13010180/everton-deducted-10-points-by-the-premier-league.

"Financial Reports - West Ham United FC." Whufc.Com, 2024, https://www.whufc.com/en/financial-reports.

"Football Transfers, Rumours, Market Values, News and Statistics." _, 2024, https://www.transfermarkt.co.uk.

Franck, Egon. "Financial Fair Play in European Club Football: What Is It All About?" International Journal of Sport Finance, vol. 9, no. 3, Aug. 2014, pp. 193–217, https://doi.org/10.1177/155862351400900302.

Government Digital Service. "Get Information About a Company." GOV.UK, 3 July 2012, https://www.gov.uk/get-information-about-a-company.

Huby, Bart, and Aaryaman Banerji. "Financial Sustainability of Men'S Football Clubs: The Premier League De-Couples from the Pyramid." Lcp.Com, 18 Dec. 2025, https://www.lcp.com/en/insights/in-brief/financial-sustainability-of-mens-football-clubs-the-premier-league-de-couples-from-the-pyramid.

Jackson, James, and James Silverwood. "A League Made in the Economy's Image: Destabilised Stability and the English Premier League's Minsky Moment." British Politics, Palgrave Macmillan, June 2024, https://doi.org/10.1057/s41293-024-00259-0.

Latif, Suleen, and Heather Lynch. “Everton’s Points Deduction Appeal: Clarity on Sanction Proportionality.” Kennedys Law, Kennedys Law LLP, 12 Mar. 2024, https://www.kennedyslaw.com/en/thought-leadership/article/2024/everton-s-points-deduction-appeal-clarity-on-sanction-proportionality/. Accessed 30 June 2026.

“Liverpool Finances 2024/25 — Swiss Ramble.” Substack.Com, 2026, https://swissramble.substack.com/p/liverpool-finances-202425.

MacQueen, J. “Some Methods for Classification and Analysis of Multivariate Observations.” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, vol. 5.1, Jan. 1967, pp. 281–98, https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Matchday Finance. “Leicester City Financial Results 2023/24.” Matchdayfinance, 27 July 2025, https://www.matchdayfinance.com/post/leicester-city-financial-results-2023-24.

---. “Premier League Financial Results 2024/25.” Matchdayfinance, 24 Apr. 2026, https://www.matchdayfinance.com/post/premier-league-financial-results-2024-25.

---. “Premier League Financial Results Season 2023/24.” Matchdayfinance, 6 May 2025, https://www.matchdayfinance.com/post/premier-league-financial-results-season-2023-24-the-results-are-in.

Müller, J. Christian, et al. “The Financial Fair Play Regulations of UEFA: An Adequate Concept to Ensure the Long-Term Viability and Sustainability of European Club Football?” International Journal of Sport Finance, vol. 7, no. 2, May 2012, pp. 117–40, https://doi.org/10.1177/155862351200700202.

Peeters, Thomas, and Stefan Szymanski. Economic Policy Fifty-Eighth Panel Meeting Supported by the Bank of Lithuania Financial Fair Play and Financial Crisis in European Football. 2013, https://www.economic-policy.org/wp-content/uploads/2013/10/peeters-szymanski.pdf.

Quinn, Paul. “The Analysis Series: The English and Major European Football Broadcast Rights Market.” Theesk.Org, 15 Aug. 2025, https://theesk.org/2025/08/15/the-analysis-series-the-english-and-major-european-football-broadcast-rights-market/.

---. “The Analysis Series: The £4 Billion Paycheck: An Analysis of Wage Cost Growth and Financial Sustainability in the Premier League (2010-2024).” Theesk.Org, 28 July 2025, https://theesk.org/2025/07/28/the-analysis-series-the-4-billion-paycheck-an-analysis-of-wage-cost-growth-and-financial-sustainability-in-the-premier-league-2010-2024/.

Ramble, Swiss. “Arsenal Finances 2023/24.” Substack.Com, The Swiss Ramble, 21 Feb. 2025, https://swissramble.substack.com/p/arsenal-finances-202324.

---. “Brentford Finances 2023/24 — Swiss Ramble.” Substack.Com, 2025, https://swissramble.substack.com/sitemap/2025. Accessed 30 June 2026.

---. “Chelsea Finances 2023/24.” Substack.Com, The Swiss Ramble, 10 Apr. 2025, https://swissramble.substack.com/p/chelsea-finances-202324.
---. “Manchester City Finances 2024/25.” Substack.Com, The Swiss Ramble, 19 Dec. 2025, https://swissramble.substack.com/p/manchester-city-finances-202425.
---. “Manchester United Finances 2024/25.” Substack.Com, The Swiss Ramble, 22 Sept. 2025, https://swissramble.substack.com/p/manchester-united-finances-202425.
---. “Premier League Debt 2024.” Substack.Com, The Swiss Ramble, 21 Aug. 2025, https://swissramble.substack.com/p/premier-league-debt-2024.
---. “Sitemap - 2024 - the Swiss Ramble.” Substack.Com, 2024, https://swissramble.substack.com/sitemap/2024. Accessed 30 June 2026.
---. “Sitemap - 2025 - the Swiss Ramble.” Substack.Com, 2025, https://swissramble.substack.com/sitemap/2025. Accessed 30 June 2026.
---. “West Ham Finances 2024/25.” Substack.Com, The Swiss Ramble, 5 Mar. 2026, https://swissramble.substack.com/p/west-ham-finances-202425.
Rambler, The Swiss. “Arsenal - Money Changes Everything.” Blogspot.Com, 2019, http://swissramble.blogspot.com/2014/09/arsenal-money-changes-everything.html. Accessed 30 June 2026.
---. “Arsenal’S Finances - 21 Questions.” Blogspot.Com, 2020, http://swissramble.blogspot.com/2011/10/arsenals-finances-21-questions.html. Accessed 30 June 2026.
---. “Show Me the Money.” Blogspot.Com, 2018, http://swissramble.blogspot.com/2013/04/show-me-money.html. Accessed 30 June 2026.
---. “The Swiss Ramble.” Blogspot.Com, 2017, https://swissramble.blogspot.com.
Rudemarlyn Urdaneta-Camacho, et al. “European Football Clubs’ Financial Performance Under UEFA Financial Fair Play: A Bibliometric Analysis and Semi-Systematic Review.” SAGE Open, vol. 15, no. 2, SAGE Publishing, Apr. 2025, https://doi.org/10.1177/21582440251340824.
Solhekol, Kaveh. “UEFA’S New Financial Sustainability Regulations to Replace FFP: All You Need to Know.” Sky Sports, 7 Apr. 2022, https://www.skysports.com/football/news/11095/12584543/uefas-new-financial-sustainability-regulations-to-replace-ffp-all-you-need-to-know.
Swiss Ramble. “English Clubs by Revenue Grouping.” Substack.Com, The Swiss Ramble, 23 Sept. 2024, https://swissramble.substack.com/p/english-clubs-by-revenue-grouping.
Szymanski, Stefan, and Tim Kuypers. Winners and Losers. Penguin, 2000.
The Unintended Consequence of Financial Fair Play: An Examination of Competitive Balance Across Five European Football Leagues. https://core.ac.uk/download/pdf/159767214.pdf.
UEFA. “Explainer: UEFA’s New Financial Sustainability Regulations.” UEFA.Com, 7 Apr. 2022, https://www.uefa.com/news-media/news/0274-14da0ce4535d-fa5b130ae9b6-1000--explainer-uefa-s-new-financial-sustainability-regulations/.

---. “Financial Sustainability.” UEFA.Com, 6 July 2023, https://www.uefa.com/running-competitions/integrity/financial-sustainability/.

“UEFA Documents.” Uefa.Com, 2025, https://documents.uefa.com/r/UEFA-Club-Licensing-and-Financial-Sustainability-Regulations-2025/Article-94-Squad-cost-rule-Online.

UEFA.com. “UEFA Executive Committee Approves New Financial Sustainability Regulations.” UEFA.Com, 7 Apr. 2022, https://www.uefa.com/news-media/news/0274-14d9f32dd3ea-b0b8ca4dca0a-1000--uefa-executive-committee-approves-new-financial-sustainabil/.

Whetstone, Sean. “The Championship & West Ham Football Finances Explained.” Blogspot.Com, 30 June 2026, http://westhamfootball.blogspot.com/2011/11/championship-football-west-ham-finances.html. Accessed 30 June 2026.

Beyond Representation: Diversity Mandates Will Not Solve Cancer Inequality By Meghna Saha

Abstract

Although cancer mortality in the United States has steadily declined in recent years, the disease still carries an enormous human cost. In 2025, researchers projected an estimated 2,041,910 new cancer cases and 618,120 deaths nationwide. Even more concerning, some minorities bear cancer mortality rates two times higher than White patients for certain cancers, such as kidney, liver, and stomach (*The Flagship Journal of the American Cancer Society*). The reason for such differences in mortality rates is partly due to representation in clinical trials. Because clinical trials determine whether new cancer treatments are safe and effective before they reach patients, the populations represented in these trials can directly influence who benefits from medical advances. However, the patients included in these trials do not always reflect the populations most affected by the disease. Some medical advances are advantageous to certain ethnicities, but not to others. It is important to remember that behind every cancer statistic is a patient whose survival may depend on treatments developed through clinical trials. In recent years, this concern has led to increased efforts to promote diversity in cancer clinical trials. Although many experts argue that mandating diversity in cancer clinical trials could improve representation and research accuracy, representation requirements alone cannot eliminate cancer disparities. Instead, the focus should be on socioeconomic barriers, healthcare access gaps, and funding instability.

Introduction

Some researchers argue that mandating diverse representation in cancer clinical trials is necessary because clinical trials directly determine which treatments are effective for different cases, such as genetic predispositions. Clinical trials are also used to make sure treatments are approved and can be used for patients. However, minority populations have historically been underrepresented in these studies, which raises concerns about how broadly the results can be applied. According to research cited in *The New England Journal of Medicine*, Black and Hispanic patients together make up less than 10 percent of participants in many cancer clinical trials, despite representing a much larger share of the overall patient population. When certain groups are absent from studies, researchers may not fully understand how treatments affect different populations. This pattern of exclusion has been widely documented. As Shriya Veluri et al. explain in *HPHR*, "[Most] racial, ethnic, and socioeconomic minority groups have been historically underrepresented in medical clinical trials," limiting the ability to apply research findings across populations. If the patients included in a study do not reflect the broader population, doctors may not fully understand how treatments will work for different groups. From this perspective, requiring diverse representation could improve the reliability and fairness of cancer research. While this argument highlights an important weakness in the current system,

representation mandates alone cannot solve the deeper structural barriers that prevent many patients from participating in trials.

Representation and Disparities in Cancer Outcomes

In addition to concerns about research accuracy, many medical experts argue that increasing diversity in clinical trials is necessary to address long-standing disparities in cancer outcomes. Without doing so, these gaps will continue to grow in healthcare. Researchers such as Meaghan Mormann et al., in *Pediatric Annals*, explain that disparities exist in diagnosis, treatment, complications, curability, survivorship, and, especially, clinical trial participation. While White patients have seen continued survival increases, patients of color have seen slower improvements. Especially Black patients whose "outcomes remain poorest." These disparities suggest that current research and treatment systems may not be fully addressing the needs of all populations. When certain groups are underrepresented in clinical trials, researchers may lack important information about how treatments affect those patients, including differences in genetics, environmental factors, or access to care. Because clinical trials help determine which treatments become standard care, some experts believe increasing diversity within them is necessary to better understand how therapies affect different populations. Similarly, according to Andrea D Carver in *Campbell University Doctor of Health Science Capstone Projects*, the lack of diverse participation undermines the scientific validity of research because it prevents researchers from accurately evaluating treatment responses across groups. In other words, when certain populations are missing from trials, researchers may overlook how treatments work differently for those groups, which can further contribute to unequal health outcomes. As a result, many experts believe that mandating diversity in clinical trials is a necessary step towards making cancer research more equitable and representative.

Supporters of diversity mandates also argue that representation is especially important because cancer itself does not affect all populations equally. Certain ethnic groups experience significantly higher rates of specific cancers and often face worse outcomes, raising concerns about whether treatments are being tested on the populations most impacted by the disease. According to data reported by *Statista* (see Fig. 1), prostate cancer occurred at a rate of 194.8 cases per 100,000 among Black men compared to only 64.5 cases among Asian American and Pacific Islander men, demonstrating how dramatically cancer incidence can vary across ethnic groups. When some groups experience much higher disease rates, it becomes especially important that they are represented in the research used to develop treatments. Researchers have also documented disparities in mortality. Elizabeth Kaplan and Anu Dairkee, in the *American Journal of Law & Medicine*, report that "Black women are 30% more likely to develop and 60% more likely to die from breast cancer than non-Hispanic White women," and they are also 40% more likely to die from breast cancer than White women. These statistics suggest that the populations facing the highest disease burden may not be benefiting equally from current medical advances. Even in cases where biological differences in tumors are not present, survival

gaps still exist. Meaghna Mormann et al. note that in hepatoblastoma, there are no differences in tumor characteristics by race, yet disparities in survival outcomes remain. This indicates that factors beyond biology, such as differences in access to care, early detection, or treatment quality, may influence outcomes. Differences can even appear within clinical trial results themselves. In the MONALESSA-7 trial reported in *The New England Journal of Medicine*, overall survival at 42 months was 83.8% among Asian patients compared with 71.5% among non-Asian patients. Variations like these suggest that treatments may not affect every population in the same way. For this reason, diverse representation is necessary to fully understand how therapies work across patient groups. Because cancer incidence and survival outcomes vary so widely across populations, supporters argue that diverse representation in clinical trials is necessary to ensure treatments are tested on the groups most affected by the disease.

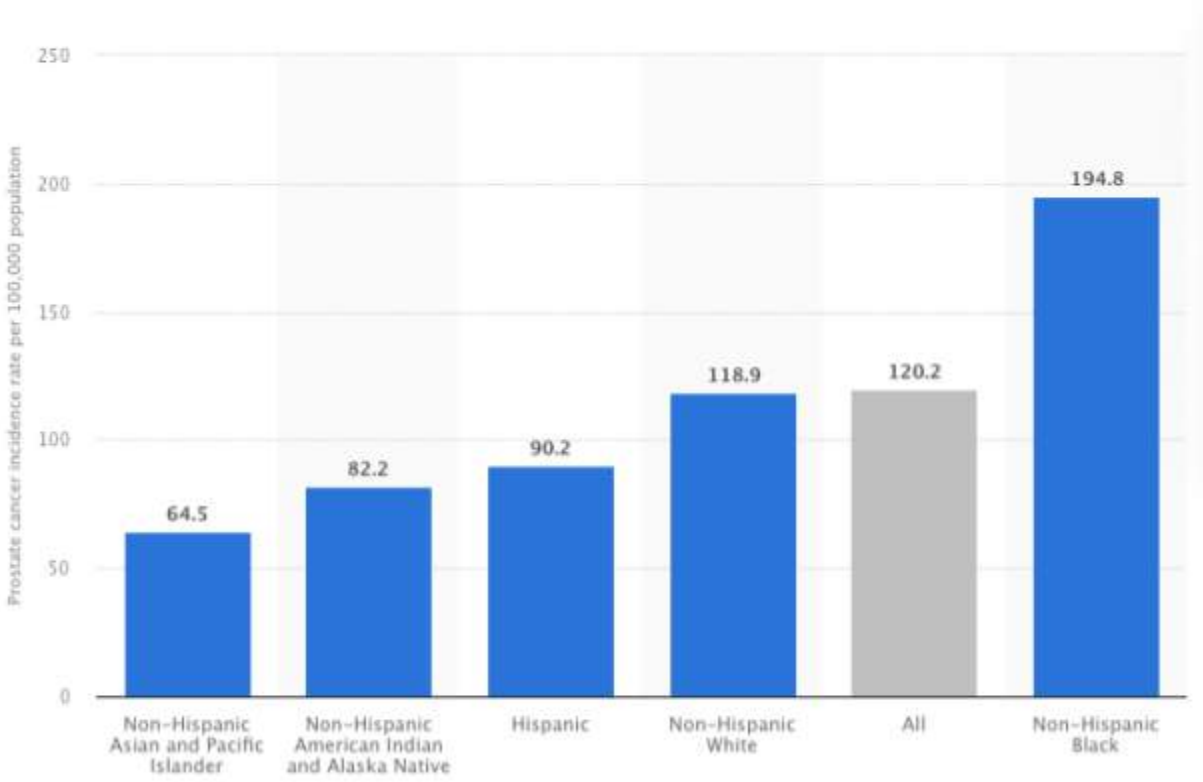


Fig. 1. Comparing the incidence rate of prostate cancer in the U.S. from 2018 to 2022 by race and ethnicity. "Prostate Cancer Incidence Rate U.S. By Ethnicity." *Statista*, www.statista.com/statistics/672946/prostate-cancer-incidence-rate-us-by-ethnicity/. Accessed on Feb 24. 2026.

Socioeconomic and Structural Barriers

Although increasing diversity in clinical trials may improve representation within research, socioeconomic barriers often prevent many patients from participating in the first place. This is especially true for low-income groups or those with no income at all, causing these groups to remain underrepresented in cancer clinical trials. Carver explains that practical challenges such as transportation costs, lost wages, and inflexible work schedules discourage participation among low-income African Americans, who remain significantly underrepresented in cancer clinical trials compared to other groups. Clinical trials frequently require repeated hospital visits, documentation, and long-term follow-up appointments, which can be difficult for patients who cannot afford to miss work or travel long distances. As a result, even when patients are eligible for trials, logistical challenges may prevent them from enrolling or completing the study. Financial barriers also affect access to treatments even when they are available. For example, *The New England Journal of Medicine* reports that "the cost of 3 years of ribociclib treatment in the United States is approximately $500,000." Researchers estimated that treating 30 patients to save one life would cost "approximately $15 million." These extreme costs make it difficult for many patients, especially those from lower-income communities, to continue treatment outside of clinical trials. These economic pressures make it far more difficult for disadvantaged populations to access the same treatments and research opportunities as wealthier patients. As a result, even if representation requirements were mandated, financial and logistical barriers would still prevent many disadvantaged populations from participating in research.

Health disparities often begin long before patients are ever considered for clinical trials, largely because of unequal access to preventive care. Preventive care refers to "[routine] health care that includes screenings, check-ups, and patient counseling to prevent illnesses, disease, or other health problems" (Kaplan and Dairkee). These services are especially important for cancer because early detection significantly improves survival outcomes. However, financial barriers frequently prevent individuals from accessing these services. Studies show that even small out-of-pocket costs can discourage patients from seeking care. In fact, Kaplan and Dairkee report that 38% of adults surveyed would not pay out-of-pocket for cancer screenings, meaning many people may delay or avoid early detection entirely. Avoiding these screenings often leads to later diagnoses, which drastically reduces survival rates. For instance, lung cancer survival is 63% when detected at an early stage, but only about 8% when diagnosed late, while cervical cancer survival drops from 91% in early stages to only 19% in late stages. These differences show how strongly early detection affects patient outcomes. Researchers also note that these access gaps continue to affect younger populations, as "persistent health disparities prevent many children and adolescents…from reaping equitable benefit from this progress" (Mormann et al.). Because many patients lack early screenings or timely diagnoses, disparities in cancer outcomes often develop before individuals ever reach advanced treatments such as clinical trials. As a result, simply increasing representation in trials cannot address the deeper structural barriers that shape who is able to access care in the first place.

Funding Instability and Research Impact

Another major limitation of relying only on diversity mandates is the instability of research funding, which can interfere with clinical trials and prevent patients from accessing them altogether. While funding has majorly been restored, clinical trials require long-term financial support to maintain research staff, recruit participants, and continue monitoring patient outcomes. When funding is reduced or delayed, these systems quickly break down, slowing research and limiting opportunities for patients who may depend on trials as treatment options. In an interview with Christiane Amanpour for *VIQ Solutions Inc.*, former National Institutes of Health director Francis Collins described the real consequences of these disruptions, giving the example of a patient with stage four colorectal cancer missing enrollment in a clinical trial because staff reductions caused delays in the research process. For patients with advanced cancers, missing an opportunity to enroll in a trial can mean losing access to potentially life-saving therapies. Funding cuts can also affect large numbers of research participants across the country. During an interview reported by *NPR*, Massachusetts Governor Maura Healey cited evidence from a 2025 study in *JAMA Internal Medicine* showing that 74,000 clinical study participants were affected by NIH funding cuts, demonstrating how quickly financial instability can disrupt ongoing research. These interruptions not only impact patients but also threaten years of scientific progress. Physician-researcher Dr. Kari Nadeau described how her NIH-funded study of 800 infants at high risk for severe food allergies was abruptly halted, putting five years of collected data at risk of being lost if funding is not restored. Situations like these show that clinical trials depend on stable infrastructure and consistent funding to continue producing meaningful results. Without reliable financial support, even well-designed studies cannot operate effectively, regardless of how diverse their participants are. As a result, addressing funding instability is essential if clinical research is to expand access and reduce health disparities.

Conclusion

Mandating diverse ethnic representation in cancer clinical trials is an important step toward improving fairness and accuracy in medical research. Increasing representation in health ensures that treatments are tested on populations that better reflect the patients who will ultimately receive them. However, representation requirements alone cannot eliminate the deeper causes of cancer disparities. Socioeconomic barriers, limited access to preventive care, and instability in research funding continue to prevent many patients from participating in trials or receiving early treatment. Because these structural challenges exist throughout the healthcare system, simply mandating diversity in trials will not fully solve the problem. To meaningfully reduce cancer disparities, policymakers must address these broader issues alongside efforts to improve representation in clinical research. Reducing financial barriers to treatment, expanding access to preventive care, and ensuring stable funding for long-term medical research are essential steps toward achieving this goal.

Works Cited

Amanpour, Christiane. Interview with National Institutes of Health Former Director Francis Collins; Interview with "2000 Meters to Andriivka" Director Mstyslav Chernov; Interview with "The Salmon Cannon and the Levitating Frog" Author Carly Ann York. Aired 1-2p ET. VIQ Solutions Inc, New York, 2025. ProQuest, https://www.proquest.com/other-sources/interview-with-national-institutes-health-former/docview/3233531778/se-2. Accessed on Feb 23. 2026.

Amanpour, Christiane, and Walter Isaacson. Interview with Harvard University Professor of Government Steven Levitsky; Interview with Harvard University Physician and Professor Dr. Kari Nadeau; Interview with "the Phoenician Scheme" Director Wes Anderson; Interview with Representative Blake Moore (R-UT). Aired 1-2p ET. VIQ Solutions Inc, New York, 2025. ProQuest, https://www.proquest.com/other-sources/interview-with-harvard-university-professor/docview/3213954255/se-2. Accessed on Feb 23. 2026.

Carver, Andrea D. The Under-Representation of Low-Income African Americans in Clinical Studies: Analysis of Clinical Trial Participation and Access to Premiere Health Care. 2024. Campbell University Doctor of Health Sciences Capstone Projects. Campbell University, JSTOR, https://jstor.org/stable/community.39164198. Accessed 26 Feb. 2026.

“Federal Appeals Court Hears Challenge to Trump Administration's NIH Funding Cuts.” NPR, Washington, D.C., 2026. ProQuest, https://www.proquest.com/other-sources/federal-appeals-court-hears-challenge-trump/docview/3292033627/se-2. Accessed on Feb 23. 2026.

Kaplan, Elizabeth, and Anu Dairkee. “The Broken Link: Braidwood, the United States Preventive Services Task Force (USPSTF), and the Health Equity Implications of Losing Free Access to Preventive Care.” American Journal of Law & Medicine, vol. 50, no. 1–2, 1 July 2024, pp. 100–120. ProQuest, www.proquest.com/docview/3150087749, doi:https://doi.org/10.1017/amj.2024.18. Accessed on Feb 25.

Mormann, Meaghan, et al. "Barriers to Equitable Pediatric Oncologic Care in the United States." Pediatric Annals, vol. 54, no. 12, 2025, pp. e426-e430. ProQuest, https://www.proquest.com/scholarly-journals/barriers-equitable-pediatric-oncologic-care/docview/3278266876/se-2, doi:https://doi.org/10.3928/19382359-20250909-02. Accessed on Feb 25. 2026.

“Prostate Cancer Incidence Rate U.S. By Ethnicity.” Statista, www.statista.com/statistics/672946/prostate-cancer-incidence-rate-us-by-ethnicity/. Accessed on Feb 24. 2026.

"Ribociclib Plus Endocrine Therapy in Early Breast Cancer." The New England Journal of Medicine, vol. 390, no. 23, 2024, pp. 2219-2222. ProQuest, https://www.proquest.com/scholarly-journals/ribociclib-plus-endocrine-therapy-early-bre

ast/docview/3069988129/se-2, doi:https://doi.org/10.1056/NEJMc2404917. Accessed on Feb 25.

Siegel, Rebecca L., et al. “Cancer Statistics, 2025.” CA: A Cancer Journal for Clinicians, vol. 75, no. 1, 16 Jan. 2025, acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21871, https://doi.org/10.3322/caac.21871.

Veluri, Shriya, et al. “Creating Space For Diversity in Informed Consent Forms.” HPHR, vol. 84, 2024. JSTOR, https://www.jstor.org/stable/48847264. Accessed 25 Feb. 2026.

Voter Turnout: How Gen Z is Impacting Climate Policy By Carla Mansilla

Abstract

My research will explore what ways this new generation of activists can influence climate policies and if a voter turnout can create changes in climate agreements. In this time the younger generation has proven they are a huge force in the fight against climate change. From organizing protests, to youth delegates, as well as their presence on social media, it has shaped to become real conversations on changing laws and public opinion. For example Greta Thunberg, the Swedish climate activist created the movement Fridays for Future.

In turn this proved to lawmakers as well as world leaders climate change was becoming a youth led issue. Using all this information on activism that the youth have employed, it still may not be enough to change policies. This is where the youth voter turnout comes into play. Thus, that is why it is important that more people should vote to prove they care. In turn, I utilized a case study analysis to examine the policies being debated in the election that inspired youth voters to turn out. My results show that Gen Z has the ability to influence the conversation on climate change for years to come. Next, voter turnout amongst Gen-Z voters shows that electoral support for the Democratic party remains high. Lastly, Gen-Z voters are more susceptible to climate anxiety. As such, the influence the youth activists voices had shows that their voices do matter and were heard.

Introduction

Global climate change is one of the biggest challenges in people's lives, and Generation Z is experiencing real consequences of the effects of climate change. The next generations are receiving this planet that is being destroyed by carbon emissions, differing sea levels, and seasons that last longer (NASA). For example, scholars concluded that more than 45% of Gen-Z experience climate anxiety which is feelings of distress about climate change that directly affects their mental health. (Hickman et.al). However, this fear is also inspiring them to care about their planet and bring awareness. Gen-Z has had more interest in environmental policies given the current administration's stance on climate change and that is why they are the most politically active generation because they demonstrated their ability at a young age to want to change their reality for the better (Livingston).

Generation Z (Gen-Z) has proven they are a huge force in the fight against climate change. From organizing protests, to youth delegates, as well as their presence on social media, gen-Z has exemplified the ability to change laws and public opinion (Chenoweth and Cebul). For example, Greta Thunberg, the Swedish climate activist paved the way in the world of youth impact. At age 15 she protested outside the Swedish Parliament advocating for change and was the first to hold a “School Strike for Climate.” Leading into the global movement Fridays for Futures that brought more than a million people to attend these protests covering 150 states (Delmestri and Schuessler). She fights to hold leaders accountable and demands them to make decisions that will help the next generations.

Furthermore, results show that Gen-Z has the ability to bring attention to policies on climate change through platforms such as Instagram, TikTok, X, and Facebook (Morris). Additionally voting participation by younger generations in 2020 helped elect Joe Biden who reentered the Paris agreement, which strives to have countries convert to a cleaner energy source and cut emissions (Davenport and Friedman). Which proves the influence the youth activists have because their voices do matter and can change anything they put their minds to. As such I argue that Gen-Z can bring attention to climate change through social media platforms as well as through civic engagement activities such as voting and organized protests. In turn this leads me to my research question; In what ways can Gen-Z influence climate policies and can voter turnout create changes in climate agreements?

As such, my research will explore what ways Gen-Z can influence climate policies and if voter turnout can create changes in climate agreements. This proved to lawmakers as well as world leaders climate change was becoming a youth led issue. Using all this information on activism that the youth have employed, it still may not be enough to change policies. This is where the youth voter turnout comes into play. That's why it's important that more people should vote to prove they care to change legislation. This leads to serious implications since Gen-Z demonstrated their ability at a young age to be engaged in politics as evidence to start new conversations about lowering the voting age population to 16.

History of Climate Change

The historical development of climate change is useful to understand why climate change has become a major political issue that influences Gen-Z voters. Climate change can often be discussed as if it is a more recent discovery, but in contrast scientific research shows that the understanding theories behind climate change have been developing for over a century (Little). It dates back to the nineteenth century when scientists were finding out that humans were directly affecting Earth's global temperature (Turrentine).This scientific history is significant because it shows that global warming is an issue that's been actively getting ignored by the government for a prolonged time.

In 1859 John Tyndall discovered the greenhouse effect this was one of the earliest breakthroughs in climate science; he found carbon dioxide and methane, which are gases in Earth's atmosphere, trap heat and regulate the planet's temperature (NASA). While this is a natural process, Tyndall proved that an excess of these gases can cause harmful global warming (Spencer). This discovery was important because it showed that an increase in gasses can lead to serious consequences and lasting harm on our atmosphere like rising sea levels, drought, and differing weather conditions.

In 1896 Svante Arrhenius expanded on John Tyndall's work by becoming the first scientist to be able to predict how an increase in CO2 increases the temperature of earth (Hofstrand). Arrhenius predicted that burning fossil fuels like coal would warm the planet (Baum). His calculations were based on limited data because of the time, nowadays modern scientists have proven his predictions were accurate. This proves that scientists were able to predict long term consequences and his work supports the argument that climate change is preventable. That could also make political passivity difficult to justify.

By 1938 climate change was no longer a concern, Guy Callendar's research showed that Earth's temperature was rising due to burning fossil fuels (UK Research and Innovation). He argued that it would continue rising if fossil fuel emissions were not reduced. His findings were significant because they connected his data to earlier predictions. His warnings show today's climate conditions, like the rising of global temperatures and changing weather patterns. Next, the strongest evidence of human caused climate change was shown in 1958 by Charles David Keeling (American Chemical Society). Keeling developed instruments to measure atmospheric carbon dioxide levels and he started by collecting data at the Mauna Loa Observatory in Hawaii (NOAA). He created the Keeling Curve which is a graph that shows a rise in CO2 levels over time (NOAA). The Keeling Curve is powerful because it is a visual demonstration of the relationship of human activity and the increasing greenhouse gases.

In the United States, the Environmental Protection Agency was created to protect human health, and ensure clean land, water, and air, for all Americans (EPA). Al Gore was a leading advocate to bring public recognition and implement policies to try to stop climate change, he helped make climate change a political topic rather than research that only interested scientists (Gore). This is important because it shows climate change is something governments should address. In the next section I will discuss voter turnout and Gen-Z politics.

Voter Turnout and Gen-Z Politics

Generation Z were born from 1995-2012 and it is the generation that grew up with the internet and social media (Florin Răduț). They are set apart from other generations because they have unlimited access to the internet and social media. This has influenced how they have learned and think about climate change. The Socialization theory, which is how people develop their beliefs that are shaped by family, friends, school, and social media explains why Gen Z cares about environmental issues (Tolstikova et al.). For Gen Z social media platforms are agents of socialization because it exposed them to climate change activism at a young age.

Climate change is Gen Z's priorities, and this can be shown in voter turnout. For example, in the 2020 presidential election, 50% of young people voted in the 2020 presidential election, which is an 11-point increase from 2016 (Circle). This turnout shows that voting is a powerful way to stop issues since it is affecting their future as well as impacting the younger generations, which makes political participation necessary rather than optional. Political psychology shows that Gen Z's increased political involvement will have lasting effects (Geser-Stark). For example, younger people are less likely to identify as a member of a political party and they describe themselves as independents while they are typically favoring one of the two major political parties (Gomez and Daniller 2021). It shows that their beliefs will shape elections and policies for decades to come (Gomez and Daniller). This shows that Gen-Z is more liberal and is more likely to vote for the democratic party (Reilly and Hunting). They take more actions towards stopping climate change. In the last section I will discuss political polarization and how Gen-Z has navigated this climate.

Political Polarization and Gen-Z's Navigation of a Divided Political Climate Political polarization in the United States has increased and climate change is one of the most divided issues between the political parties (Layman, Carsey, and Horowitz). There is ongoing growing

ideological conflict between political parties most defined by the Democrats and Republicans. Democrats view climate change as an important issue that can be stopped if more people advocate for it and Republicans view climate change as a lower issue on the global scale tending to focus more on the economy related issues. This clear divide shapes how climate change is discussed in politics and also influences how Gen Z feels about political information. There is mainly loyalty over beliefs which in turn causes activism. Next, I will discuss how social media serves as a platform for Gen Z to remain politically active.

Social media is a huge activist platform. Platforms such as TikTok, Instagram, X, Facebook are the places Gen Z goes to get their information (McClain et al.). These platforms allow Gen Z to share information, and organize movements (McClain et al.). However echo chambers are created in those platforms and feelings of distrust, rebellion, and anger are created (Travers). This can increase polarization but it also motivates activism and overall political engagement. Despite polarization, Gen Z has shown a strong ability to promote change around global warming. Activists create Grassroot movements that link to justice and equality. An example of this is Fridays for Future and activists like Greta Thunberg (Engelfried). Through protests, online campaigns, and voting, Gen Z fights for climate change to be an important political issue (Gomez). Gen-Z will shape the future of our government with their strong activism and ability to bring social influence to all aspects. When Gen-Z reaches their voting age their abilities will grow and define the future political landscape

Discussion and Conclusion

Throughout my research I prove that Gen Z is an active and significant force in determining what the political aftermath will be of the ongoing climate change discussion. Gen Z has continued to demonstrate their activism by gaining support through the digital world, social campaigning, and overall political participation. In fact, this shows a shift from previous generations and their traditional policies (Carnegie). Therefore, my discussion of the history of climate change is evident that the government has ignored global warming, especially because this is not a new controversy. This inaction to save our planet has grown Gen Z's passion, urgency, and sense of responsibility to be more environmentally and politically engaged than the previous generations (Leyts).

My most significant findings include the growing relationship between voter turnout and youth activism that Gen Z demonstrates. While Gen Z advocates through different channels like digital activism, they do put pressure on policymakers, however it does not always turn into a law or reform (Steinberg). Additionally, their activism pays off with an increased voter participation that political leaders have to acknowledge. An example of this is the 2020 presidential election when former president Biden reentered the Paris Climate Agreement (Casler et al.). Considering the high rate of youth turnout that voted for a political leader whose beliefs matched their climate policies, this proposes that Gen Z has significant influence when they are civically engaged.

For example, Gen Z has utilized social media to increase their political engagement. TikTok, Instagram, and X create opportunities for activists to dissect various levels of

information, and make climate change an issue that is morally related to justice. This research also shows the limitations of what social media activism can do, for example echo chambers. Echo chambers create political polarization that strengthen beliefs that continue to split the parties apart (Kitchens et al.). Political polarization creates conflict as well as contributing to Gen Z's identity that promotes a civic engagement for climate change.

Climate change is considered to be one of the most polarizing political divides in American politics which brings challenges and struggles for Gen Z (Eccles). Gen Z has shown their capability of navigating around these divided political situations by promoting what the global consequences are and overall long term impact on our world. From staying away from a specific party, their independence grants them the option to give their attention to policy outcomes. Gen Z is a force that continues to change how climate change discussions are being recognized in future elections.

In conclusion Gen Z has proven to the world that their voices are important and indispensable when it comes to climate change as well as showing climate change can be a youth navigated issue. Their continuous political engagement is challenging the traditional norms of policy issues, and is making sure that global warming is one of the main debates being had. When more Gen Z reach the legal voting age their power over the political landscape will completely redefine and guarantee that policies will have everyone's best interests due to the climate change crisis. A long term implication of Gen Z's political perseverance proposes that the central concern in politics will be climate change because it is being discussed by a generation that focuses on our long term stability and survival.

Works Cited

American Chemical Society National Historic Chemical Landmarks. “The Keeling Curve.” Baum. *(you referenced Baum earlier but it’s missing full citation—fix this)*

BillionBricks. “Gen Z at the Forefront of Climate Advocacy - BillionBricks | Net-Zero Homes.” BillionBricks, 19 May 2023.

Carnegie, Megan. “Gen Z: How Young People Are Changing Activism.” *BBC*, 8 Aug. 2022. Casler, Don, et al. *Do Pledges Bind? The Mass Politics of International Climate Targets.* 2025. Chenoweth, Erica. “Why Gen-Z Is Rising – Ash Center.” *Ash Center*, 13 Jan. 2026. CIRCLE. “Election Week 2020: Youth Vote May Be Decisive in Key Senate Races.” 25 Nov. 2020.

Dajana, Muhammad Hassan, and Syeda Hamna Shujat. “See How Youth Worldwide Are Now Influencing Climate Action.” *World Economic Forum*, 12 July 2023.

Davenport, Coral, and Lisa Friedman. “Biden, in a Burst of Climate Orders, Rejoins the Paris Agreement.” *The New York Times*, 20 Jan. 2021.

Delmestri, Giuseppe, and Elke S. Schuessler. “The Mission (Im)Possible of Climate Action through Quixotic Institutional Work.” *Journal of Management Studies*, 2025.

Eccles, Robert G. “How Polarized Is America When It Comes to Climate Change?” *Forbes*, 5 Feb. 2024. Emma. “The Role of Youth Activism in Shaping Political Change.” *Medium*, 24 Aug. 2023.

Engelfried, Nick. “How Generation Z Is Leading the Climate Movement.” *Waging Nonviolence*, 14 Jan. 2020.

EPA. “Our Mission and What We Do.” *United States Environmental Protection Agency*, 1 May 2024. Florin Răduț. “Generation ‘Z’ and Social Networks.” *Journal of Educational Studies*, 2021. Gen. “Generation ZALPHA.” 17 Aug. 2024.

Gomez, Melissa. “Young Climate Activists Find Powerful Voice in Social Media.” *Los Angeles Times*, 12 Sept. 2024.

Gore, Al. “Climate Leadership.”

Hofstrand, Don. “The Greenhouse Effect Is Proven Science.” 2021.

Hughes, Susan A. “Young Voters Shifted Right in the 2024 Election…” *Harvard*, 2025.

Kitchens, Brent, et al. “Understanding Echo Chambers and Filter Bubbles…” *MIS Quarterly*, 2020.

Layman, G. C., Carsey, T. M., & Horowitz, J. M. “Party Polarization in American Politics…” *Annual Review of Political Science*, 2006.

Little, Becky. “Industrial Revolution Linked to Early Signs of Climate Change.” *HISTORY*, 2023. Livingston, Kelly. “Gen Z, Worried about Climate Change…” *ABC News*, 2024. NASA. “The Effects of Climate Change.” 2024.

NASA Earth Observatory. “John Tyndall (1820–1893).”

NOAA. “Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases.” 2024.

Reilly, Thom, and Dan Hunting. “Understanding the Gen Z Independent Voter…” *Politics*

& Policy, 2025.
Steinberg, Stacey. “#Advocacy: Social Media Activism’s Power to Transform Law.” 2017. Tolstikova, Irina, et al. “Generation Z Social Capital as a Result of Digital Socialization.” 2021. Travers, Mark. “How to Break out of the Echo Chamber.” *Psychology Today*, 2023. Turrentine, Jeff. “What Are the Causes of Climate Change?” *NRDC*, 2022.
UK Research and Innovation. “A Brief History of Climate Change Discoveries.” 2021. Weart, Spencer. “The Carbon Dioxide Greenhouse Effect.” 2019.

Breathing Breaks During Lecture-Style Classes: Impacts on High School Students' Short-Term Comprehension of New Astronomy Content By Shreya Bollapally

Introduction

In high school classrooms, students are expected to maintain focus throughout long lectures, especially in subjects that are very content-based, such as science courses. In these classes, students process large amounts of new information at once, while teachers attempt to sustain their students' attention during classes (Müller et al., 2021). Prior research demonstrates that students need a positive and engaging learning environment in order for them to develop a growth mindset, which is essential for academic success (Nayyar et al., 2025). Similarly, researchers Peiris et al. (2021) claim that students can find it difficult to focus in a setting with little engagement, and that they need to learn in an environment that works best for them. In response, teachers have started to implement various classroom interventions in order to support their students' learning and performance. One common strategy that teachers use is implementing short and structured breaks during class. There are many ways these breaks can be designed, including physical activity breaks, breathing breaks, or stretching breaks. In a study by researchers Campbell and Lassiter (2020), they claim that structured breaks may support students' attention. However, different types of breaks can have various impacts on aspects of student learning and performance.

A specific type of classroom break is using breathing breaks, which researcher Norouzi (2019) describes as short, intentional pauses involving mindful breathing. The benefits of breathing breaks, such as improvement in overall well-being, are based on mindfulness, which researchers Müller et al. define as "paying attention in a particular way: on purpose, in the present moment, and non-judgementally" (Müller et al., 2021, p.2). Breathing-based breaks are accessible to be used in classrooms because most methods do not require additional resources or physical movement, making them feasible to implement in classrooms. These factors make structured breathing breaks a practical and potentially beneficial intervention to support high school students in their learning.

Based on these previous studies that have examined the impacts of break interventions in classrooms, high school students may benefit from short breaks to allow them to fully process and absorb the new information they are learning. This topic is especially relevant to high school students, who are often tested on this content-heavy material based on their comprehension and understanding of the subject and what they have learned. In order for students to perform better in schools, where they are assessed regularly on the content they learned, it is important to study how to improve lesson comprehension (King, 1991). This heavy emphasis on academic performance is especially prevalent at the high school level. Even though there is existing research demonstrating that mindfulness breathing can be beneficial for high school students to help their attention and well-being, there is a lack of research on whether structured breathing breaks can improve short-term comprehension during a lecture-style lesson.

These factors have led me to address the following research question: Do high school students demonstrate greater short-term comprehension of new astronomy lecture material when short breathing breaks are included?

Literature Review

In order to understand how students learn in a classroom, it is important to look at and understand many factors that connect to and impact each other, including attention, comprehension, instructional material, and the types of strategies that are used to support students' learning. When exploring the existing scholarly conversation, many researchers have already studied the impacts of structured breaks in a classroom setting on various aspects of student learning. Many studies reference a common foundational text by John Sweller as the theoretical basis for explaining why these breaks are actually beneficial. These advantages are explained through Cognitive Load Theory (Sweller, 1988). In this theory, Sweller (1988), a frequently cited and well-known researcher in cognitive psychology, explains that working memory can only hold a certain amount of information, which means that students can only process a certain amount of information at once. In a lecture-style lesson where students can be presented with a lot of information at once, they can experience cognitive overload, which can impact their overall comprehension of the lecture material. Sweller's theory helps explain that breaks might be beneficial because they interrupt continuous lectures and give students a chance to mentally recover. Cognitive Load Theory is used by many researchers, making it a credible concept to understand the impacts of classroom break interventions. This idea is supported by other researchers, such as Norouzi (2019), whose research focuses on the impacts of microbreaks, who claims that breaks may reduce cognitive fatigue. Additionally, in another study that explores the impact of breaks during a task, the researchers claim that breaks may result in "an improvement of learning and retention" (Capeda et al., 2009; Donovan & Radosevich, 1999, as cited in Blasche et al., 2018). These studies come together to support the idea that interrupting a long lesson can support learning for students. This exploration of existing scholarly literature about classroom breaks supports the idea that they can be beneficial to students and might improve their ability to process new information during lecture-style classes.

According to these sources, many researchers have claimed that structured breaks can support how well students process new information and their overall retention and attention in class. In order to justify my choice of studying short-term comprehension, it is essential to identify a connection between attention and comprehension. It is important to make the distinction between attention and comprehension because a student may appear more attentive or focused, but that does not mean they understood the material more effectively. In a recent study, researchers Bergey et al. (2023) found that attention predicts student performance, and students who stay focused in class are more likely to perform better in school. Similarly, in a study by researcher King (1991), who studies educational psychology, he demonstrates that

attention improves comprehension and emphasizes the importance of focus when understanding new material. The relationship between attention and comprehension helps identify a connection between how attention directly supports comprehension, which then supports student performance in a class. These studies suggest that implementing classroom breaks, which have been previously found to improve attention, may therefore improve students' comprehension, aligning with my exploration of breathing interventions on students' short-term comprehension.

One method of working to support students' cognitive abilities and attention in classrooms is implementing breathing interventions, which are part of the more broad category of mindfulness. Breathing interventions are a strong and practical choice to explore as a classroom intervention for various reasons. Researchers Campbell and Lassiter (2020), who focus on barriers to implementing classroom interventions in their study, explain that many teachers are hesitant to incorporate physical activity breaks into their lessons because they require space, and they often disrupt the classroom schedule. However, compared to physical activity breaks, breathing interventions are less disruptive and more feasible to add in classroom settings because they do not require movement or any extra resources. Similarly, in a study by researcher Sellakumar (2015), he claims that breathing interventions are feasible to implement in classrooms. However, just feasibility is not enough to justify why breathing breaks can be beneficial. A strong argument comes from researcher Norouzi (2019), who explains the benefits based on cognitive functioning and claims that breathing breaks can be a potentially strong classroom intervention because they have been found to improve attention and school performance (Norouzi, 2019). These are factors that directly correlate with short-term comprehension.

Even though other researchers have established that there is a correlation between breaks, attention, and comprehension, the existing scholarly literature does not fully examine how short breathing interventions during a lecture-style lesson impact immediate short-term comprehension. Studies such as Norouzi (2019) and Sellakumar (2015) suggest that breathing activities can support students' attention, reduce stress, and improve different aspects of their performance. Additionally, researchers such as Müller et al. (2021), whose research examines various classroom interventions, study the impacts of classroom interventions on long-term student performance and connect mindfulness practices to improved performance, rather than short-term comprehension outcomes. This previous research provides support that there is value in breathing interventions in classrooms, but there is a gap in the literature about if these interventions lead to improvements in students' short-term comprehension. This gap is significant when looking at the participants: high school students who are often expected to understand new material quickly during lecture-style lessons and are evaluated about their comprehension through tests and quizzes. If this specific intervention is found to be supportive for students, it can be a feasible idea to implement in classrooms that can benefit the students' academic performance (Han, 2025).

In order to study this gap, it is important to choose instructional material that allows for an accurate measurement of students' short-term comprehension after a lecture-style class. Using astronomy as a lecture topic matches this goal. In an article by researcher Sadler (1992), he claims that astronomy is a concept-heavy subject that is often unfamiliar to high school students. To further support this lack of knowledge among students, in a more recent study on astronomy education by researcher Percy (2024), he explains that many undergraduate students are not familiar with astronomy concepts as well. Because high school students typically do not have strong knowledge in astronomy, their performance in the study would be more likely to reflect what they learn from the lecture, rather than what they already know.

The sources that exist in the scholarly conversation suggest that there is a relationship between attention, breathing interventions, and comprehension. However, there is a gap in the literature about how short breathing interventions in a lecture impact short-term comprehension, which is especially relevant to high school students. Studying this gap would lead to a better understanding of the impact of short breathing interventions and can provide a potential classroom strategy to benefit high school students.

Method

This study used an experimental design with a control and experimental group to understand whether short breathing breaks influence high school students' short-term comprehension of new instructional material. This design was selected because it allows two groups to be compared: one receiving an intervention and one that does not. This method is especially useful to isolate and understand the impact of the independent variable on the dependent variable, which is the impact of short breathing interventions on students' short-term comprehension in this case. Researchers Voltmer et al. (2023) used a similar method in their study, in which participants were randomly assigned to a control and experimental group in order to isolate the impacts of the independent variable that is different between the groups.

Participants

This study involves 22 high school students who were recruited through a voluntary sign-up Google form. Students from grades 9 to 12 were allowed to participate; however, the majority of students who participated were juniors and seniors, with less representation of freshman and sophomore students. This group of participants was chosen because high school students can potentially benefit from the implications of this study, as they often experience lessons that are lecture-based and are assessed on their comprehension of them through tests and quizzes. The participants who signed up were randomly assigned to one of two session dates, which were then assigned to be either the experimental or control group. This process resulted in there being 11 participants per group.

Materials

The instructional materials are made up of three educational videos from the *Crash Course Astronomy* series: *Introduction to Astronomy: Crash Course Astronomy #1, Naked Eye Observations: Crash Course Astronomy #2, and Cycles in the Sky: Crash Course Astronomy #3,* with a total duration of approximately 35 minutes (Crash Course, 2015a, 2015b, 2015c). High school students are often not familiar with many astronomy concepts (Sadler, 1992). These videos were selected because they provide introductory information about astronomy, which ensures that the participants' scores will reflect what they learned through the lecture and will reduce prior knowledge bias in the study. The lecture style of the videos and the overall duration are also meant to reflect a lecture in a real classroom setting.

Comprehension was measured using a Google Form pre-test and post-test with 12 multiple-choice questions each based on the content of the videos. The tests were created by me based on the videos and are not externally validated. The full set of pre-test and post-test questions is included in the Appendix. Since there were 3 videos, the questions were split up so questions would correspond with each video. Even though the tests were not externally validated, the questions are accurate and aligned with the content presented in the lecture videos. The same questions were used for both groups in both the pre-test and post-test to make sure that the measurement was consistent. The use of a pre-test and post-test set-up is supported by prior research, such as a study by Zhou et al. (2025) in which they used this design to measure changes in student learning after a spaced learning intervention. The pre-test and post-test structure was chosen over another type of evaluation because it allows for a baseline testing of the participants' initial knowledge of astronomy.

Procedure

Both groups went through a mostly similar procedure. In each separate group, the participants took a pre-test to understand their baseline knowledge of basic astronomy concepts. They then watched the instructional videos, and they immediately took an identical post-test to measure their understanding after the video. The participants were not informed that the tests would be the same to reduce testing bias and to make sure that score improvements were due to comprehension rather than memorizing the questions. The participants were also not allowed to see their pre-test scores or use any other devices during the study, so they would not already know the answers before taking the post-test.

The control group watched the videos continuously without any breaks. The experimental group watched the same videos but participated in two structured breathing breaks during the session. Each break was approximately 1-2 minutes and followed the 4-7-8 breathing method, which involves inhaling for 4 seconds, holding for 7 seconds, and exhaling for 8 seconds (Priasmoro et al., 2025). This is a common breathing method found to reduce fatigue and be beneficial for students and their academic performance (Priasmoro et al., 2025). The procedure was designed in a way to balance breathing breaks with the continuation of the lecture-style lesson to make sure that the breaks did not significantly interrupt the lesson and

distract the students. This procedure aligns with real classroom limitations because teachers have to make sure to balance schedule and timing limits, along with keeping their students engaged.

Three participants in the control group completed the study in a different location due to scheduling conflicts after already signing up to participate in the study. Rather than having uneven sample groups of 11 and 8, I decided to set up an alternative testing date for these 3 participants in an attempt to align with the original procedure and maintain consistency. This location was arranged to closely simulate the original testing environment at the school. I attempted to maintain very similar testing conditions in the different locations, such as similar lighting, similar video set-up, and few distractions.

Results

A total of 22 students participated in this study, with 11 students assigned to the experimental group and 11 students assigned to the control group. All of the participants completed a 12-question multiple-choice pretest before they watched the lecture video and an identical 12-question post-test immediately after the lecture video. Each question is worth 1 point, so the maximum score for participants was 12 points. To understand the effect of breathing breaks on short-term comprehension, I began by calculating the gain scores for each participant by subtracting the pre-test scores from the post-test scores. I conducted a two-tailed paired-sample t-test for each group to understand if there was a significant change in scores from before the videos to after. I also conducted an independent samples t-test to compare the gain scores between the experimental and control groups.

Based on the average gain scores for all participants, both groups improved from pre-test to post-test. The experimental group had a mean gain score of 5.00 points, while the control group had a mean gain score of 4.55 points (see Figure 1 below).

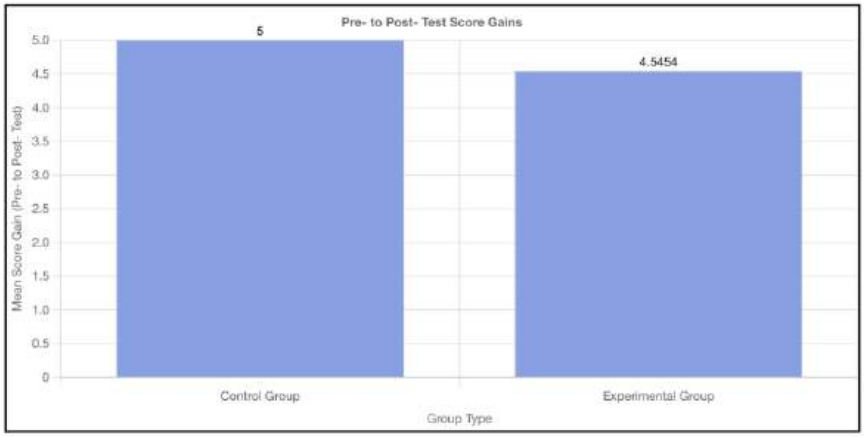


Figure 1

Comparison Within Groups

In order to understand if participants in each group demonstrated statistically significant improvement, I conducted two two-tailed paired-sample t-tests for the control group and experimental group. This test was appropriate because each of the participants had two scores that were related: a pre-test and a post-test. The test would help me compare changes with the same participants before and after the lecture. A two-tailed test was specifically used because I could identify if there were any statistically significant differences in the scores, instead of assuming that the scores of all participants improved in the post-test.
For the experimental group, I began by calculating the t-value, which represents the ratio between the sample mean differences between the test and the standard error of the differences. I found that the t-value is 9.92, and the degrees of freedom, which is based on the number of pairs, is 10. With these conditions, in the experimental group, $p < 0.001$. The p-value indicates how likely it is that the improvement happened because of random variation. In this case, since the p-value is less than 0.05, the result is statistically significant, and the improvement is not likely to be because of random variation.

Similarly, in the control group, I followed the same steps. The increase in the control group is also statistically significant, as the t-value is 8.11 and the degrees of freedom are also 10. As a result, $p < 0.001$, which means that the scores of students in the control group also improved from the pre-test to the post-test.

Comparison Between Groups

In order to identify if the breathing intervention contributed to a greater improvement in scores, I conducted an independent samples t-test to compare the gain scores of participants in the control group and the experimental group. This test is appropriate for the situation because both groups had different participants, which makes the samples independent. Based on the difference between the mean gain scores between both groups, the participants in the experimental group improved by 0.45 points more on average than the participants in the control group. However, according to the independent samples t-test, this difference is not statistically significant. I calculated that the degrees of freedom are 20 and the t-value is 0.60. This t-value is small compared to the variation in each group. Based on this result, the p-value is greater than 0.05, which means that the slight improvement in score could have been because of random variation, rather than as a result of the breathing interventions.

According to the statistical analysis, the participants in both groups had a statistically significant improvement in short-term comprehension based on the pre-test and post-test scores. However, the difference in gain scores between the two groups was not statistically significant. Even though both groups improved on their own, the slightly higher improvement of the experimental group can not be attributed to the breathing intervention. These results mean that the findings could have been a result of random variation, and there is not enough evidence to identify a clear impact of the breathing interventions.

Conclusion

Teachers continue to search for educational strategies to improve student performance and engagement in classes. This issue is especially prevalent in classes that are lecture-based and content-heavy, such as many science classes, in which students are expected to process a large amount of new information. Research that already exists suggests that classroom breaks and interventions based on mindfulness can support students' attention and classroom performance (Sellakumar, 2015). However, there is less research that focuses specifically on whether short breathing interventions during lecture-style classes improve short-term comprehension.

In this study, I researched whether high school students demonstrated greater short-term comprehension of new astronomy lecture material when short breathing breaks were included. The results demonstrated that the scores of participants in both the control and experimental groups improved significantly from the pre-test to the post-test, which means that the students did learn from the lecture-style video. However, the slight increase in improvement of the experimental group compared to the control group is not statistically significant. As a result, according to this study, there is not enough evidence that is statistically significant to prove that structured breathing interventions in a lecture-style lesson improve short-term comprehension in high school students.

These results are significant because they provide a qualifier for the general scholarly understanding that classroom breaks can lead to better academic performance from students. These results support and also add to the existing scholarly conversation. Other researchers, such as Norouzi (2019) and Sellakumar (2015), claim that there are benefits to mindfulness practices on attention and overall performance, but these results highlight that they might not directly correlate with short-term comprehension. However, the results of this study demonstrate that this understanding might be more complex, and there might be some factors, such as short-term comprehension, that are not influenced by mindfulness breaks. Even though breathing breaks may still support attention and focus, the results of this study show that they may not be effective enough on their own to lead to an increase in short-term comprehension. These findings contribute to the existing scholarly conversation by giving evidence to a more specific gap about the impacts of breathing breaks on short-term comprehension after a lecture-style class.

There are some limitations to consider when interpreting and understanding these conclusions. One of the main limitations is that the sample size was small, with only 11 participants in each group. This limitation reduced statistical power and made it more difficult to identify smaller differences and trends between the groups. Since the sample size is small, the lack of statistically significant evidence might not fully represent the impacts of the breathing intervention because there is less data to analyze. This limitation is significant because the experimental group had a slightly higher mean gain score, but the sample size might not have been big enough to figure out if that difference was statistically significant. Similarly, Müller et al. (2021) reported a similar limitation in their study about classroom breaks and explained how

the difference in size of the groups can impact the conclusions from their data. Another limitation is that the sample group does not fully represent all high school students. Although students from grades 9 through 12 were allowed to participate in the study, the participants were not evenly split. Most participants who signed up were juniors and seniors, with only a few freshmen and sophomores, which limits how well I can generalize these findings to all high school students, especially younger students who might have had a different reaction to the lecture-style videos. Additionally, the pre-test and post-test to measure short-term comprehension were created by the researcher and are not externally validated. Even though the questions are based on the content in the video, and they are standardized for all participants, they may not be a fully accurate method to track short-term comprehension. A final limitation is that three participants in the control group completed the study in a different testing environment on a different day due to scheduling conflicts. Although I tried to maintain similar conditions in the testing environments, the participants might have been slightly influenced by a change in some small factors, which could have impacted their overall comprehension of the video and, therefore, their scores on the tests.

Even though there are some limitations of the conclusions, the results of this study do have implications in the real-world. These findings can be beneficial to all high school teachers who are looking for classroom interventions to support high school students. Short breathing breaks can be easy to implement in classrooms and support student attention, which is demonstrated in a study by researcher Norouzi (2019), as well as by other researchers. However, according to the results of this study, they might not improve short-term comprehension alone and could be used along with other classroom interventions. This result is important because many teachers might assume that breathing interventions can improve academic performance, but their results might actually depend more on the context and specific details of the type of intervention. This idea provides opportunities for future research. Future researchers can study how short breathing interventions used with other classroom strategies, such as note-taking methods or lighting variations, can improve short-term comprehension. Additionally, researchers can study the impacts of various types of breathing methods used during these short and structured breathing breaks. These slight changes can contribute further to the existing scholarly conversation about types of classroom interventions that can be successful in improving short-term comprehension and provide teachers with information about how to best support their high school students' academic performance.

This study is important because it contributes to scholarly conversation on the impacts of a specific type of breathing intervention on short-term comprehension and provides a direction for future research to explore similar types of feasible classroom interventions to support students.

Appendix

Pre-Test and Post-Test Questions

1. How did studying the sky help ancient humans?
2. Which early belief claimed that Earth was at the center of the universe? 3. What invention helped astronomers capture faint objects before digital detectors existed? 4. How much of the universe can we see directly?
5. What is it called when you observe the sky with no binoculars, telescope, or other instruments?
6. What system ranks stars by brightness?
7. What problem makes faint stars harder to see in cities?
8. What are sets or groups of stars in the sky called?
9. Why do different constellations become visible at different times of the year?
10. What is the name of the line the Sun appears to follow through the sky? 11. What causes the seasons on Earth?
12. What is the slow wobble of Earth's axis called?

Works Cited

Bergey, B. W., Cromley, J. G., Kaplan, A., & Bloxton, J. D. (2023). Do students' questions during chemistry lectures predict perceived comprehension and exam performance? The Journal of Experimental Education, 91(3), 411-430. https://doi.org/10.1080/00220973.2021843

Blasche, G., Szabo, B., Wagner-Menghin, M., Ekmekcioglu, C., & Gollner, E. (2018). Comparison of rest-break interventions during a mentally demanding task. Stress and Health, 34(5), 629-638. https://doi.org/10.1002/smi.2830

Campbell, A. L., & Lassiter, J. W. (2020). Teacher perceptions of facilitators and barriers to implementing classroom physical activity breaks. The Journal of Educational Research, 113(2), 108-119. https://doi.org/10.1080/00220671.2020.1752613

CrashCourse. (2015a, January 15). *Introduction to astronomy: Crash Course astronomy #1* [Video]. YouTube. https://youtu.be/0rHUDWjR5gg

CrashCourse. (2015b, January 22). *Naked eye observations: Crash Course astronomy #2* [Video]. YouTube. https://youtu.be/L-Wtlev6suc

CrashCourse. (2015c, January 29). *Cycles in the sky: Crash Course astronomy #3* [Video]. YouTube. https://youtu.be/01QWC-rZcfE

Han, J. (2025). Mindfulness mediates the relationship between academic stress and academic achievement among college students. Social Behavior and Personality: An International Journal, 53(5), 1-8. https://doi.org/10.2224/sbp.14236

King, A. (1991). Improving lecture comprehension: Effects of a metacognitive strategy. Applied Cognitive Psychology, 5(4), 331-346. https://doi.org/10.1002/acp.2350050404

Müller, C., Otto, B., Swaitzki, V., Kanagalingam, P., Scherer, J. -S., & Lindberg, S. (2021). Short breaks at school: Effects of a physical activity and a mindfulness intervention on children's attention, reading comprehension, and self-esteem. Trends in Neuroscience and Education, 25, 100160. https://doi.org/10.1016/j.time.2021.100160

Nayyar, S., Verma, S., Kalidoss, D. K., Yadav, S., Kashyap, B., & Sharma, S. (2025). Educational psychology and learning styles: Adapting teaching methods for student success. Journal of Neonatal Surgery, 14(19S), 969-977.

Norouzi, A. (2019). *Employee one-minute mindful breathing microbreaks for present-moment attention, recovery, and well-being* [Doctoral dissertation, Grand Canyon University]. ProQuest Dissertation & Theses.

https://www.proquest.com/openview/153f85310c8181e6c3b8e9b4e73cf4b3/1 Peiris, C. L., O'Donoghue, G., Rippon, L., Meyers, D., Hahne, A., De Noronha, M., Lynch, J., & Hanson, L. C. (2021). Classroom movement breaks reduce sedentary behavior and increase concentration, alertness and enjoyment during university classes: A mixed-methods feasibility study. International Journal of Environmental Research and Public Health, 18(11), 5589. https://doi.org/10.3390/ijerph18115589

Percy, J. R. (2024). Reflections on undergraduate astronomy education. Journal of the

Royal Astronomical Society of Canada, 118(5), 224-226.
Priasmoro, D. P., Asri, Y., & Maulina, R. (2025). Exploring 4-7-8 breathing for stress relief and improved quality of life in chronic and degenerative diseases: A scoping review. *Proceeding International Conference of Innovation, 5*(1), 209-217.
Sadler, P. M. (1992). The initial knowledge of high school astronomy students. Harvard Smithsonian Center for Astrophysics.
Sellakumar, G. K. (2015). Effect of slow-deep breathing exercise to reduce anxiety among adolescent school students in a selected higher secondary school in Coimbatore, India. *Journal of Psychological and Educational Research, 23,* 54-72.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4
Zhou, Y., Hartley, R., Bernardelli, A., & Tolmie, A. (2025). The impact of spaced learning within physics lessons in secondary schools. PLOS ONE, 20(4).
https://doi.org/10.1371/journal.pone.0321552

The Capital Gains Tax: Its Impact on Socioeconomic Disparity in 21st Century America By Christopher Page

Wealth inequality has become one of the defining features of the American economy in the twenty-first century, and few policy tools illustrate this trend more clearly than the capital gains tax. Defined by the IRS as a tax on the profit realized when a capital asset is sold, the capital gains tax was designed to treat investment income as a legitimate, taxable source of wealth. Yet the way this tax has been structured, and more importantly, the loopholes that have grown around it, has allowed the wealthiest Americans to shield enormous sums of income from taxation while ordinary wage earners continue to pay a proportionally higher rate on their labor. This raises an urgent question: to what extent has the capital gains tax impacted socioeconomic disparity in the United States thus far in the twenty-first century, and how can policymakers use it to bridge the financial gap between the wealthy and lower classes? Throughout the twenty-first century, socioeconomic disparity has increased in the United States, and this is in part due to loopholes in the capital gains tax. To close these loopholes, the United States should raise the capital gains tax rate and begin taxing unrealized gains as well.

Before examining the evidence, it is worth clarifying the scope of this argument. The analysis that follows focuses specifically on the United States rather than attempting cross-national comparisons. Other countries operate under fundamentally different tax systems, economic structures, political institutions, and government policies, all of which shape how a capital gains tax functions and what effects it produces. Attempting to compare how such taxes affect inequality in nations with no capital gains tax, a different tax structure, or different governing systems altogether would introduce too many confounding variables. Narrowing the scope to the United States allows for a clearer, more isolated view of how this specific tax interacts with domestic wealth inequality.

The current state of the capital gains tax reveals a striking imbalance. Capital gains are typically taxed fifteen to seventeen percentage points lower than ordinary income in the United States, meaning that a teacher or nurse earning a salary is taxed more heavily on that income than a wealthy investor is taxed on profits from selling stock. Meanwhile, the top one percent of Americans holds an estimated forty-four percent of all unrealized capital gains, and the top one percent's share of total national wealth grew from twenty percent to twenty-seven percent over the course of the twenty-first century (Montecino et al., Inequality.org, 2025). Research from the Equitable Growth research group underscores just how concentrated this form of income truly is: from 2002 to 2021, the top one percent of the income distribution received 45.3 percent of all capital gains, while the top ten percent received nearly three-quarters, or 75.7 percent (Robbins et al., Equitable Growth, 2025). Capital gains have also come to represent roughly twenty percent of national income over the past two decades, and in 2021 alone, aggregate capital gains reached $16.2 trillion. Remarkably, less than thirteen percent of these gains were ever reported on income tax returns, meaning the vast majority of this wealth largely escapes taxation altogether.

This pattern of avoidance is not accidental; it is built into the structure of the tax code itself through a series of well-established loopholes. Provisions such as the so-called "Angel of Death" loophole, primary residence exclusions, 1031 like-kind exchanges, tax-advantaged retirement accounts, and qualified small business stock exemptions all allow wealthy taxpayers to defer, reduce, or entirely eliminate their capital gains tax liability (Fox & Liscow, 2026). Perhaps the most emblematic of these strategies is the practice known as "buy, borrow, die." Under this approach, a wealthy individual buys appreciating assets, borrows against them using low-interest, tax-free loans to fund their lifestyle rather than selling the assets outright, and then passes those assets on to heirs at death, at which point the accumulated gains are wiped clean for tax purposes under a stepped-up basis. This strategy allows enormous fortunes to compound and transfer across generations while contributing little to no tax revenue along the way, deepening the very disparity this essay seeks to address.

The consequences of allowing these gains to remain largely tax-free are significant. When investment income becomes effectively tax-free relative to wages, wealth compounds more rapidly among those who are already wealthy, widening the gap between asset owners and everyone else. Maintaining and strengthening the capital gains tax is one of the few available checks on this compounding effect; without it, the disparity would materialize even faster than it already has.

Efforts to weaken the capital gains tax have not gone unchallenged politically. When President Trump attempted to eliminate the tax, the effort failed to pass in Congress, in part because analysts warned that abolishing it would further widen the racial wealth gap, given how unevenly capital assets are distributed across racial groups (Gravely, Center for Public Integrity, 2018). This political scrutiny reflects a broader recognition that the capital gains tax, whatever its flaws, plays an important role in the nation's approach to wealth distribution.

Raising the capital gains tax, rather than weakening it, offers several economic benefits beyond simply narrowing the wealth gap. Higher capital gains revenue could help reduce federal deficits and debt, which in turn helps maintain stable interest rates. Contrary to the assumption that higher taxes on investment discourage economic activity, research suggests that higher capital gains tax rates would be mildly expansionary and could reverse the "crowding-out" effect that government debt has on private investment (Foster et al., American University, 2025). Freed from the distortion of preferential tax treatment, capital could also flow more readily toward business investment, job creation, and wage growth, potentially encouraging businesses to expand into low-income areas and improving economic conditions in underserved communities across the country. This finding directly challenges the common assumption that lower rates on capital gains are necessary to spur growth; the evidence instead suggests that eliminating tax preferences for capital gains income would reduce inequality, correct inefficiencies in the tax system, and raise significant government revenue all at once.

Taken together, the evidence points to a clear conclusion: socioeconomic disparity in the twenty-first century United States has been driven, at least in part, by a capital gains tax system riddled with loopholes that disproportionately benefit the wealthiest Americans. The solution is

not to abolish the tax, as some have proposed, but to strengthen it, closing loopholes like "buy, borrow, die," raising the rate on high-income households, and extending taxation to unrealized gains. Doing so would not only generate meaningful public revenue and support economic stability, but would also begin to close the widening financial gap between America's wealthy class and everyone else.

Works Cited

IRS. (2025, January 2). Topic No. 409 Capital Gains and Losses | Internal Revenue Service. Irs.gov. https://www.irs.gov/taxtopics/tc409

Monroe, M. (2025, February 26). New research finds capital gains are highly concentrated and hardly taxed, underscoring widespread U.S. inequality. Equitable Growth. https://equitablegrowth.org/new-research-finds-capital-gains-are-highly-concentrated-and-hardly-taxed-underscoring-widespread-u-s-inequality/

Raising Capital Gains Taxes Would Reduce Inequality Without Economic Costs. (2025). Inequality.org. https://inequality.org/article/capital-gains-taxes-inequality/

Gravely, A. (2018, August 21). Capital gains cut would widen racial wealth gap. Center for Public Integrity. https://publicintegrity.org/inequality-poverty-opportunity/capital-gains-cut-would-widen-racial-wealth-gap/

The Rich's Real Tax Trick Isn't "Buy, Borrow, Die." (2026, June 15). Tax Policy Center. https://taxpolicycenter.org/taxvox/richs-real-tax-trick-isnt-buy-borrow-die

Foster, M., Gonzalez Garcia, I., Hansen, M., Montecino, J., & Mooney, W. (2025). Technical Note on the Economics of Taxing Capital Gains (Version 4). Institute for Macroeconomic and Policy Analysis. https://doi.org/10.57912/28404686

The Impact of Song Lyrics on Emotional Response: A Multidisciplinary Perspective By Qianyi Wang

Abstract

Music has the unique ability to evoke a wide range of feelings, from elation and happiness to sadness and longing. While the instrumental nature of music clearly contributes to its emotional impact, song lyrics are a dynamic medium of communication that can enhance and enrich listeners' emotional responses. Whether by virtue of storytelling, evocative imagery, or common theme, song lyrics can profoundly affect listeners, invoking intense emotional responses that are culture- and language-independent. Therefore, it is not just theoretically significant to understand how song lyrics provoke feelings but also pragmatically relevant to a variety of disciplines ranging from music therapy to advertising and media communication. This paper explores the role of song lyrics on emotional response from a multidisciplinary perspective, informed by evidence from psychology, neuroscience, linguistics, and musicology.

Introduction

Music is often referred to as a universal language owing to its exceptional ability to convey emotions and concepts transcending to cultures (Mehr et al., 2018). Whereas melodies set the mood for a song, words provide an instant and intimate mode of expression that can become very powerful for the listeners. The power of song lyrics lies in the ability to express sophisticated emotions, convey interesting stories, and connect the gap between the artist and listener. While passing on happiness, sorrow, love, or anger, lyrics deepen emotional experience, and as a result, they are an indispensable part of the impact of a song (Brattico et al., 2011). The best thing about song lyrics, arguably, is the ability they have to provoke rich emotional reactions. Similar to poetry, song lyrics use imagery, metaphor, and storytelling to build emotional scenery in which one can identify as an individual (Barradas & Sakka 2021). For instance, one might become emotionally reduced to tears through Adele's ballads, yet Pharrell Williams' "Happy" will immediately brighten up one's day. Music even has the capacity to evoke a feeling we didn't know we had in the first place, hence making lyrics such a unique stimulus for both emotional self-discovery and catharsis.

There are a few factors that render the lyrics emotionally compelling such as subject matter, the language itself, and the manner in which the song is delivered (Stewart & McAlpin, 2015, see Table 1). The songs based on common themes such as love, heartbreak, yearning, or courage have a tendency to establish a deeper emotional rapport between the artist and the listener. Music is best able to describe human experience in language other people have trouble articulating themselves, so the listener can relate their own feelings more to the song's message. Even the way words are put on paper influences the way emotion can be provoked. Authors will incorporate sensory details and figurative language at times in an effort to create mental pictures that convey more sense of emotional significance in a song. Descriptive narratives enable listeners to “see” emotions, places, or periods that make the music more personal and

meaningful. Beyond the literal lyrics, the way that lyrics are written influences the listener's emotional response as well (Levy et al., 2024). Whether through passionate belting or sensitive exposure, the tone, pitch, and emotional interpretation of a singer can add meaning to a song. Intensive tone creates richness of feeling, while figurative language infuses richness and allows listeners to interpret emotion in a personal way. These processes working together transform words into expressions of emotion as well as regulators of the extent to which emotions are felt.

Table 1. Key elements of song lyrics and their emotional impact on listeners

Element of Lyrics	How It Influences Emotion	Example
Relatable Themes	Connects to personal experiences and memories, making listeners feel understood.	Love, heartbreak, resilience (e.g., 'Someone Like You' by Adele)
Figurative Language	Enhances emotional depth by using metaphors and similes to convey complex feelings.	Metaphors (e.g., 'Firework' by Katy Perry)
Imagery and Symbolism	Creates vivid mental pictures that intensify the listener's emotional engagement.	Visual storytelling (e.g., 'Boulevard of Broken Dreams' by Green Day)
Repetition and Rhyme	Increases memorability and emotional impact by reinforcing key lyrical messages.	Repeated phrases (e.g., 'Shake It Off' by Taylor Swift)
Delivery and Performance	Amplifies the emotional weight of lyrics through vocal expression and dynamics.	Passionate vs. subdued vocals (e.g., 'Rolling in the Deep' by Adele)
Genre and Style	Shapes the listener's interpretation based on expectations associated with music genres.	Rock vs. pop vs. folk influences (e.g., 'Bohemian Rhapsody' vs. 'Shape of You')
Cultural and Personal Context	Allows for deeper emotional connections when lyrics align with cultural or personal identity.	Culturally significant lyrics (e.g., protest songs, national anthems)
Language and Word Choice	The choice of words, tone, and phrasing influences the emotional tone of the song.	Poetic vs. conversational tone (e.g., 'All Too Well' by Taylor Swift vs. rap lyrics)

Apart from evoking feelings, song lyrics also serve as an effective means of emotional release and expression (Barradas & Sakka 2021). Songwriting gives the artist a medium through which to explain their own personal lives in words, turning raw emotion into art. For listeners,

music offers a catharsis, which allows them to release their emotions. It finds solace in words that it can identify with, expressing feelings that it cannot voice. Therefore, music is both an emotional escapism and a means of self-exploration. It is an instrument that helps individuals navigate their experiences and thoughts.

Since teenagers are constantly exposed to music, song lyrics are particularly important for them. Besides, adolescence is a period of identity formation and emotional development. For young people, music can become a means to accept emotions and bond with other people (Chen, 2023). Songs can also empower and console them during crises. This review of literature explores the emotional impact of song lyrics from a comprehensive perspective, integrating findings from various disciplines. By taking into account how lyrics influence emotional experience, this paper attempts to explain the profound ways in which music builds our thought and way of life with emphasis on youth listeners.

Psychological Mechanisms in Song Lyrics' Emotional Impact

Lyrics in songs possess a unique ability to impact emotional responses on the basis of a series of interconnected psychological processes (DeWall et al., 2011). At the center of this ability lies the semantic meaning of lyrics, which conveys feelings and ideas that listeners can internalize. When song lyrics speak to universally human themes such as love, loss, perseverance, or self-discovery, they are most likely to strike a strong chord with individuals, provoking emotional reflection and a feeling of unity between listener and artist. For young audiences like high school students, these themes are even more so (Sánchez González 2021), as adolescence is a time of intense emotional experience and self-definition. The tune of a song, whether about self-esteem or personal struggle, can resonate like a personal expression of one's own feelings. This makes music a highly personal and compelling experience.

One of the strongest psychological influences behind song lyrics is the ability to trigger autobiographical memory. It has been shown that music is a powerful cue to memory, evoking past memories along with their linked emotions in startling richness. A study by Janata et al. (2007) found that music-evoked memories correlate with heightened levels of emotion and can persist over extended periods of years. The experience, known as music-evoked autobiographical memory, occurs when the lyrics of a song resonate with a listener's experience or past events. For example, a student who has experienced a painful breakup can find solace in a ballad that reflects the feelings of the singer, while a perseverance song can be used as motivation in school hardship. It is common that there are certain songs in the high school years, such as a prom or a marching song, which become deeply associated with shared memories and recall emotions after years of not having heard them.

The emotional power of song lyrics is also increased by the use of literary devices, which add layers of meaning to the lyrics. They consolidate abstract emotions into tangible expressions. Metaphors, in particular, enable listeners to more readily relate to complex emotions by framing them in familiar imagery. A good example is Leonard Cohen's "Hallelujah," where the recurring metaphor of a "broken hallelujah" evocatively expresses the simultaneity of joy and despair.

Through the use of figurative language, songwriters can create lyrics that are personal but broadly relatable. The lyrical nature of song lyrics makes them more emotionally powerful because they allow for personal interpretation. When people hear metaphors being applied in a song, they do not just process them literally. Instead, they relate them to their own experiences and make the song feel deeply relevant to their own lives. Consequently, the emotions conveyed in the song become more evocative, as each person connects with the lyrics in a unique and meaningful way.

The pattern of a song's narrative also has a vital function to convey feelings. Most songs are narratives and follow an introduction, followed by the middle part, and then resolution. This structure allows for emotional escalation, towards points of climax. In Adele's "Someone Like You," for instance, the building of emotion incrementally through the verses and chorus to a crescendo of emotional release is, for many listeners, cathartic. Such formal trajectory models emotional trajectories in lived experience, and listeners have no trouble becoming engaged with the emotional material of the song.

The second of the main mechanisms of emotional response to lyrics is emotional contagion: the process whereby listeners are mirroring the artist's emotional state. Lauria (2023) describes how music, as with human contact, can "infect" listeners with affective states. When a singer complains with honest sadness, her delivery in lyrics and tone of voice have the effect of inducing sadness from listeners who have not personally suffered the event to which the singer is relating within the song. This has the effect of requesting empathy, because listeners are able to empathize emotionally with the singer's perspective as a temporary part of her life. High school teenagers, whose social awareness and empathy are still developing, may be particularly influenced by the lyrics of songs treating such issues as bullying or social injustice. A song like "Stand By You" could prompt listeners to reflect on their own friends or think about a person who requires their support, which deepens their emotional connection to the song.

The pairing of music and lyrics is also significant for emotional impact. Vuoskoski and Eerola (2012) confirmed that music with lyrics was perceived as more emotionally evocative than instrumental music alone. This study provides strong evidence that words give more emotional background information to add to the listener's sense of connection to the music at a deeper level. They facilitate more straightforward expression of feelings so that listeners are better able to discern the song's intended mood. Lyrics also have the capacity to stimulate autobiographical memory and personal association, as discussed earlier. This finding is in agreement with greater studies indicating that verbal expression in music increases emotional identification and investment. Schubert (2013) demonstrated that emotional responses to music are the product of a dynamic interaction between lyrics, melody, rhythm, and vocal performance, emphasizing the importance of considering both textual and musical elements in shaping listener emotions.

Lastly, music for teenagers is far more than mere entertainment; it is an emotional companion that directs them through the turbulent pubescence (Sánchez González, 2021). Lyrics reflect personal experience in a way that makes one's emotions tangible and understandable. The raw honesty embedded in lyrics allows teenagers to be heard, enabling them to deal with emotions of

loneliness or confusion. Whether through lyrics that express their distress or tunes that motivate and challenge, music is a safe emotional release for self-regulation. By choosing songs that share their current emotions, teens utilize music to sort through their emotions and build a sense of self. It is this deep, personal connection between lyrics and emotion that makes music so integral a part of the teenage experience, shaping both their self-image and social reality.

Neurological Underpinnings of Song Lyrics' Emotional Impact

The affective impact of song lyrics goes beyond psychological processes and is firmly based in neuroscientific mechanisms. Analyzing how the brain processes lyrics uncovers a complex interaction between multiple neural systems. It has been shown that listening to music with lyrics activates a broad network of brain regions that integrate linguistic and affective processing to create a rich listening experience (Toader et al., 2023; see Figure 1 for a flowchart). How lyrics affect emotions is not only what is being said, but also how the brain processes and reacts to those words within the context of music.

When a person listens to a song, the brain employs a number of specialized areas that work together to encode both lyrical and musical aspects. The auditory cortex, located in the temporal lobe, is responsible for analyzing the basic acoustic attributes of music, such as pitch, rhythm, and tone. But when there are lyrics, other language-processing regions are involved, including Broca's area (speech production) and Wernicke's area (meaning processing). Both of these regions, located in the left hemisphere, help the listener process the words and syntax of the lyrics, distinguishing them from the strictly instrumental aspects of the song.

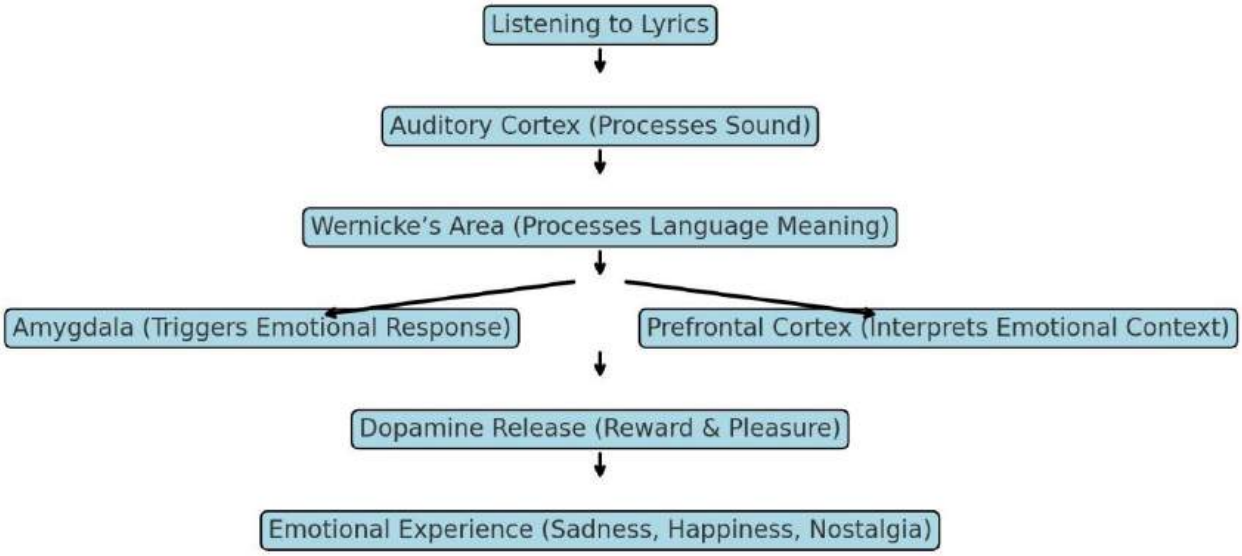


Figure 1. A flowchart showing how lyrics are processed in the brain and linked to emotions.

At the same time, the amygdala, the brain's central emotional processing center, is highly active, especially if a song evokes strong emotional experiences. Neuroscientific studies using fMRI scans (Brattico et al., 2011) have found that sad songs with lyrics are more highly activated in the

brain than instrumental sad music, most notably in the amygdala, the parahippocampal gyrus (in terms of memory and emotion), and the frontal lobe (which helps in meaning interpretation and emotional connotation). This suggests that lyrics reinforce the feeling of sadness, both stimulating the brain's cognitive and affective centers. Surprisingly, it has been found that upbeat music without words activates the limbic system of the brain (dealing with emotion processing) more so than upbeat music with words. This would indicate that lyrics are particularly well-suited for describing more complex or melancholic emotions, while mere happiness might be expressed by the cheerful tones of the music itself.

The brain integrates music and lyrics to form an integrated emotional experience. Koelsch et al. (2005) found in a study that listening to lyrics in music activated both auditory and linguistic processing areas like the superior temporal gyrus and the inferior frontal gyrus. It seems that the brain does not process lyrics separately from music but as an integrated emotional experience of emotion and language. This is the reason why lyrics will contribute more to the emotional impact of a song than they are a standalone narrative element.

Aside from activating certain parts of the brain, song lyrics trigger biochemical responses that influence mood. Music is known to release dopamine, a neurotransmitter associated with pleasure and motivation. The dopamine hit happens in the nucleus accumbens, the brain's central region of reward circuitry, and accounts for the fact that some songs are so arousing and even addictive (Ferreri et al., 2019). Emotional lyrics may be more conducive to this, particularly if individuals sing along or fully relate to the lyrics. For example, an uplifting chorus or inspiring lyrics may further increase self-esteem and motivation, once more activating the reward response.

Other neurochemistries also play roles in modulating emotional responses to lyrics. Comforting music with soothing lyrics have been shown to reduce cortisol, the primary stress hormone, and thus anxiety (Chen, 2023). Additionally, the hippocampus, a region that plays a role in forming memory, is highly active when lyrics trigger autobiographical memories. This is the reason why it is possible to connect a familiar melody with a burst of past emotions, generally leading the listener back to that specific moment in their life.

Unsurprisingly, not all individuals process music and lyrics in the same way. Variations in brain structure and function between individuals can lead to variations in emotional responses to song lyrics. Several factors, including experience and personality, shape the manner in which listeners perceive and relate to lyrics. In particular, musicians and individuals with extensive musical training are neurologically more sensitive to music and lyrics. Pantev et al. (2003) found that musicians had increased cortical representation of musical tones, meaning that they process musical elements more accurately. This enhanced auditory processing would also most certainly extend to lyrical appreciation because musicians are more sensitive to how the language plays with the melody. Their ability to discern subtleties of intonation and lyrical nuances allows them to hear music in a more emotionally differentiated way than non-musicians.

Personality and personal life experiences are also known to influence individuals' reaction to song lyrics. People who are open to experience like complex, avant-garde music with

intricate, intellectually stimulating lyrics. Similarly, individuals who are empathetic or have high emotional sensitivity react more strongly to emotional lyrics. Huron and Vuoskoski (2020) found that empathetic individuals experience stronger emotional responses to melancholic music because they engage more with the narrative and affect in the lyrics. In addition, lyrics that align with a person's personal experiences are more likely to cause strong emotional responses as they activate the brain's self-referential processing networks. Janata et al. (2007) confirmed that music-evoked autobiographical memories were associated with increased activation in the medial prefrontal cortex, a region of the brain involved in personal identity and self-reflection. This is why music linked to important life events, such as relationships or struggles, evokes strong emotional reactions.

Overall, neuroscience shows that song lyrics engage the brain in a multiplicity of ways, combining language processing, emotional interpretation, memory recall, and reward system activation into one activity. Lyrics are not alone but instead interact dynamically with the musical elements of a song to modulate the degree to which an emotion is felt and remembered. For teenage listeners this means that music becomes an effective means of self-expression and identity formation. A song can legitimize a feeling, provide comfort, and even serve as a source of inspiration, showing just how closely lyrics of music and human emotion are intertwined.

Linguistic Features of Song Lyrics and Their Emotional Impact

Undoubtedly, the linguistic features of song lyrics play a significant role in establishing their emotional impact. Aside from the literal message of words, lyrics convey emotions through a multitude of stylistic elements. Whilst rhyme gives lyrics musicality and convenience, repetition increases dominant ideas, intensifying their emotional connotation. Besides, poetic words, as metaphors and imagery, increase depth in the aspect that listeners interpret song lyrics at their own level. Vocal technique, in tone and intonation, adds emotional delivery in the aspect that the message is amplified and expressive. All these together provide an extended aural and cognitive experience that can amplify emotions and heighten memorability. Examining how these devices function will help us better understand why certain songs resonate with listeners so strongly.

Rhyme is arguably the most fundamental language feature of song lyrics. Rhyming words have a repetitive sound that is appealing to the ear and more easily remembered, and thus makes the lyrics more memorable and duplicates the emotional force of the song. Repetition of the same sounds produces an impression of shape and solution, something that may elicit emotions of satisfaction and emotional identification. It was found in research by Patel et al. (2006) that rhyming lyrics are processed more effectively in the brain than non-rhyming lyrics. This is why the latter are likely to be found more captivating, memorable, and emotionally stronger. Rhyme can also be used strategically to highlight the central emotions or subject matter of a song. In Eminem's "Lose Yourself," for instance, the tight internal rhyme scheme produces a sense of urgency and emotional intensity that reflects the song's motivational theme.

Similarly, rhythm in song lyrics determines the emotional meaning of a song by coordinating the words with the speed of the music. Lyrics to a fast rhythm can intensify excitement, urgency, or energy, while lyrics with slow, stretched-out phrasing can intensify melancholy, contemplation, or tranquility. Research by Tierney and Kraus (2013) has shown that music rhythm processing is strongly linked with social and emotional communication, with those rhythms that resonate with a listener's internal tempo giving a greater sense of emotional connection. A fast rap verse can convey self-confidence and aggression, and a slow ballad can convey sadness and longing. In younger audiences, rhythmic coordination might particularly be intense among those songs linked with social identity, such as school fight songs or prom song played on occasions of specific life events.

Another strong literary device that is used to make lyrics catchy and emotionally resonant is repetition. Repetition in words, phrases, or even entire choruses by artists is used to reinforce emotional messages so they can be recalled for a very extended period of time after the song has ended. Margulis (2013) arrived at the conclusion that repetition in music has the power of enhancing listeners' engagement and deepening emotional response because familiarity breeds attachment and emotional connection. Nunes et al. (2014) also demonstrated how songs that have extremely repetitive lyrics tend to gain more popularity, as the "repetitive priming" phenomenon makes listening to the song more pleasurable over time.

Many of the world's most recognizable songs rely heavily on repetition. For instance, in Pharrell Williams' "Happy," the repeated phrase "Because I'm happy" is not only catchy but also psychologically reinforces positive emotional states. Similarly, in Adele's "Someone Like You," the repetitive chorus underscores the themes of loss and longing, making the song more poignant. The pop music chorus is an emotional foundation, repeating the same message of the song over and over again to get a strong impact. The repetition makes the songs more memorable and easier to connect to personal experiences among young people, thereby creating a stronger relationship with the music.

Aside from structural means like repetition and rhyme, figurative language imbued with emotional depth and readability is also an elementary part of lyrical narrative. Simile, metaphor, and imagery are common ingredients songwriters use in order to transpose intangible emotions into concrete sensory experiences other people can imagine and perceive. Using metaphors, songwriters are able to parallel emotion with material events such that these feelings are more easily known and more dense in evocativeness. For example, Kendrick Lamar's "DNA" uses the DNA metaphor to explain how ingrained personal traits and struggles are. Instead of simply declaring his identity and resilience, he encodes these into a metaphor so they are more resilient and suggestive. In Taylor Swift's "Begin Again," the phrase "I watched it begin again" is utilizing simple yet powerful imagery to describe hope and rebirth following heartbreak. Imagery, or that language working with description invoking the senses, serves to strengthen the emotional hearing of a song. In "Ocean Eyes" by Billie Eilish, "burning cities and napalm skies" constitutes the image and thereby an image-driven and emotionally orienting word comparison to stimulate the emotional comprehension and visualization in listeners, for which no rational meaning

appears suggested. By imagery and their abilities, songs draw word-pictures in such descriptive words so powerful in conveys.

For adolescents who are most commonly drawn to emotional and abstract lyrics, figurative language is a means through which emotions can be conveyed indirectly so that listeners can infer and become more strongly attached emotionally. Teen listeners most commonly identify with metaphorical and lyrical lyrics because they express the incoherence of adolescence, wherein emotions are unstable and not possible to put directly into words.
How vocals convey lyrics is also critical in establishing their emotional impact. Elements like intonation, tempo, and timbre influence the manner in which emotions are conveyed and felt (See Table 2). A study by Juslin and Laukka (2003) demonstrated that these element differences significantly influence the way in which the listener perceives the emotions embedded in lyrics. It is for this reason that one can sing identical lyrics and, based on delivery, produce significantly different emotions. Leonard Cohen's original "Hallelujah" has a rich, somber timbre that inspires spirituality and contemplation, whereas Jeff Buckley's cover is more celestial, lamenting in its tone, intensifying the melancholy and the yearning of the piece.

Table 2. Impact of vocal elements on emotional perception

Element	Definition	Impact on Emotion	Example
Intonation	The rise and fall of the voice pitch that conveys different emotional states.	A rising intonation can indicate excitement or questioning, while a falling intonation suggests sadness or finality.	A singer increasing pitch at the end of a phrase can create a sense of uncertainty or anticipation.
Tempo	The speed at which lyrics are sung, affecting emotional intensity and urgency.	Fast tempos convey energy, urgency, or joy, whereas slow tempos evoke sadness, introspection, or solemnity.	A slow ballad tempo enhances feelings of heartbreak, while an uptempo song energizes and uplifts.
Timbre	The quality or tone of the voice that adds depth and texture to emotional expression.	A rich, deep timbre can make lyrics feel more somber, while a breathy or bright timbre can enhance vulnerability or hopefulness.	Leonard Cohen's 'Hallelujah' has a somber timbre, while Jeff Buckley’s cover has a celestial, lamenting tone, altering emotional perception.

Cultural and Contextual Factors in the Emotional Impact of Song Lyrics

The emotional connotation of song lyrics is not only determined by their musical and linguistic characteristics, but is also powerfully influenced by cultural values, expectations, and the context under which they are received. Personal experience, the surrounding social situation when a song is released, and cultural background affect the meaning and emotional commitment

to lyrics from the listener's perspective. These factors create a complex interdependence among music, culture, and identity, making song lyrics extremely subjective and, at times, world-defining.

Cultural values and social norms influence the meaning of song lyrics and emotional reception. Those lyrics that verify popular cultural themes and social experiences touch deepest within people of the same culture. Family-centered, communal, national, or value-moral-kind-of songs can create strong emotions within people having the same beliefs. Boer et al. (2013) conducted a study that revealed music taste and emotional reactions to a song are greatly determined by culture. The study showed that people from various cultures have varying emotional responses to a song, depending on their social values and connections. For example, a song of personal independence and freedom would create pride and happiness in individualist cultures, but the same song would generate ambivalence or even conflict in collectivist cultures if the message is not compatible with collective cultural beliefs. A fascinating example is Bruce Springsteen's "Born in the U.S.A.," which contains themes of patriotism and social conflict. While most listeners perceive it as a patriotic song, others, with closer examination of the lyrics, see it as a protest against war and the struggles of working-class Americans. This dualism in interpretation helps to demonstrate how cultural background and personal outlook influence emotional response to lyrics.

The genre and musical arrangement of a song influence the way listeners interpret its emotional message. Each genre has certain expectations for the content of lyrics, the style of vocals, and instrumental accompaniment, which shape the expression and perception of emotions. A study conducted by Rentfrow et al. (2011) found that different musical genres draw out various emotional profiles among listeners. Pop music, for instance, will often convey happiness, love, and empowerment through repetition and catchy lyrics. Blues and country music also tend to delve into heartbreak, perseverance, and nostalgia, which tend to be relatable for listeners who have endured personal adversity. Hip-hop and rap tracks tend to emphasize social struggles, identity, and empowerment, which resonate with listeners who can identify with themes of resilience and social justice. Rock music tends to express energy, revolution, and personal struggle, with powerful instrumentals serving to drive home the emotional gravity of the lyrics. Electronic and dance music lyrics are usually lean and repetitive, bringing about an aura of euphoria and exhilaration in clubbing or festival settings.

Performance style affects emotional content in lyrics too. A moving ballad could transform bland lyrics into an emotional experience, or an energetic rock performance could even render depressive lyrics angry-sounding. A good case in point is with Adele's "Someone Like You," where the strength of her voice and a slow, expressively emotional delivery heighten the potency of the song, when otherwise the same words sung quickly without much emotional belief would be disappointing.

Other than cultural background, personal preference and character also decide the way people emotionally connect with lyrics. People naturally connect with songs that reflect their emotions, values, or upbringing, leading to highly individualized reactions to the same song.

Hargreaves et al. (2006) showed that musical taste was typically found to be associated with personality traits. People with high openness to experience will prefer complex, cerebral lyrics, getting more engaged with music that challenges popular themes. People with high emotional sensitivity will be finding and deeply moved by melancholic, self-reflective lyrics, as indicated in a study by Vuoskoski & Eerola (2012). It is the listeners who have high levels of nostalgia who are most likely to listen to emotionally charged lyrics related to experiences and memories, creating an even higher emotional level. That is why two people can be listening to the same song but experiencing very different feelings, depending on their personality, past experience, and mood while listening.

Mood and Personal Experiences as Emotional Triggers in Lyrics

Obviously, the emotional effect of lyrics is influenced by the current mood and personal experience of the listener. Experiments have established that listeners relate more to those words that are closest to their immediate emotions. A study by Hunter et al. (2011) found that individuals in a sad mood tend to rate sad music as emotionally rewarding and evocative, an effect termed mood-congruency effect. This is why one going through a breakup will find comfort in heartbreak ballads, but another individual in a celebratory mood will choose happy, victorious lyrics.

Furthermore, music can trigger autobiographical memory as discussed previously. Janata et al. (2007) demonstrated how specific songs could trigger vivid memories of past events, and that lyrics are employed as affective time capsules. An example would be listening to a song from one's high school years and it being able to bring forth the feelings and memories of those years simultaneously, highlighting the emotional power of lyrics. For young people, music can be a way of processing feelings and creating individual identity. Teenagers often link lyrics to their own experiences, challenges, dreams, and relationships, as a form of self-definition and emotional coping.

It has been uncovered in studies that while some emotional cues in music are universally cross-cultural, cultural context becomes the prevailing influence for determining overall emotional reactions. Balkwill & Thompson (1999) determined that even basic emotions such as happiness or sadness in music can still be recognized across cultures, but more intricate emotional meanings heavily rely on an individual's cultural affiliation. We witness this in songs of protest during the past and present, where they carry high cultural significance and emotional connection among different audiences. For instance, Bob Dylan's "The Times They Are A-Changin'" was itself such a movement anthem for the 1960s, hitting a note within civil rights and anti-war activist groups. John Lennon's "Imagine" is a globally famous peace song, but its reception varies based on cultural and political perceptions. K-pop music, which integrates Western pop conventions with Korean cultural themes, is now widely popular worldwide, yet its lyrics are given greater emotional significance by Korean audiences who understand the cultural references and idiomatic connotations unique to the language.

For teenagers struggling to navigate identity and self-expression, genre-based music offers intense emotional release enabling them to articulate and work through emotions. Music serves as a road map of life's peaks and valleys. Take, for instance, the raw, rebellious anger of punk rock, which captures frustration and disillusionment with society, speaking to teens who feel misunderstood or constrained by others. In contrast, rap music's rhythmic aggression and direct storytelling tend to empower the listener, reinforcing themes of determination and self-worth. Meanwhile, genuine ballads provide solace in times of heartbreak, allowing listeners to grieve lost love or friendship while gaining a sense of solidarity in the lyrics. Whether by forceful guitar chords, rhythmic conversations, or soft melodic songs, the language of music presents a close emotional connection that enables adolescents to recognize their own lives in the words and rhythms that they get caught up in.

Music Therapy and Emotional Healing Through Lyrics

Music therapy is a well-established field that uses music interventions to improve emotional health. By understanding how lyrics move emotions, therapists can utilize them as a means for individuals to work through feelings and develop emotional coping mechanisms (Levy et al., 2024). Studies have shown that song lyrics help manage emotions. They give listeners a sense of validation and support. Lyrics that recreate a client's own experience can be employed as a healthy outlet for discovery, wherein individuals are able to articulate complicated feelings that may be difficult to discuss.

Therapists often use lyric analysis in sessions, where clients listen to, discuss, or even write their own lyrics to explore their emotions. Silverman (2011) investigated if music therapy combined with role-playing would be able to improve assertiveness skills in psychiatric inpatients. Despite a lack of differences in control of the environment and quality of life, those who took part in the assertiveness groups had greater internal locus of control (feeling that they could affect what occurred) and quality of life than those in the non-music condition. The most frequent attendance was reported by the music therapy group, and people liked music sessions compared to non-music assertiveness training as more enjoyable and therapeutic. These results suggest that music therapy sessions incorporating songs with meaningful lyrics significantly improved emotional expression in clients with mental health challenges. For instance, a person in mourning will enjoy a song that talks about loss and recovery. The lyrics can give them the sense that they have been heard, and they are not alone. This sense of fellowship in their pain comforts them.

Music therapy has also been found to be highly effective for veterans coping with post-traumatic stress disorder (PTSD). In a study by Bensimon et al. (2008), veterans who participated in therapy sessions that involved discussing and writing songs about their trauma reported reduced PTSD symptoms and improved emotional regulation. By verbalizing their experiences through lyrics, participants were able to process their trauma in a structured and supportive environment. The study demonstrates how music can be a therapeutic bridge between internal emotions and external expression.

For high school students, lyrics provide an accessible and relatable way to explore emotions. Research indicates that 93% of young people (ages 16–24) use song lyrics as a "therapeutic tool" to navigate mental health challenges (Stock, 2023). In addition, 73% of teen listeners recognize that reading, listening to, or writing words helps them to work through painful emotions. Teens naturally gravitate towards music when hurting, either by hearing words validating their hurt or by creating pieces themselves as expression. That would suggest music therapy sessions in teens' individualized programs specifically for teens might be a beneficial method of emotional nurturing.

Lyrics in Advertising and Media

Other than therapy, song lyrics also serve as powerful tools in advertising and marketing, where they are used to create emotional connections between consumers and brands. Advertisers deliberately utilize music with emotive lyrics to influence audience perceptions. As a result, they are able to strengthen brand recall and enhance consumer engagement without being overtly intrusive or forceful. Arguably the most successful example is the 1971 Coca-Cola “I'd Like to Buy the World a Coke” campaign's use of unity- and happiness-evoking lyrics to support the brand message (Keller & Spence, 2023). McDonald's "I'm Lovin' It" campaign also employed the use of an easy-to-sing-along-to jingle, making the phrase and thereby the brand even more recognizable upon hearing. Alpert et al. (2005) found through a study that background music which is emotionally pertinent increased purchase intent and improved brand attitudes, demonstrating how the emotional content of music impacts consumer decision directly. Lyrical use is made discreetly in advertisements to manipulate the mind of consumers. For instance, an advertisement for a cell phone based on a song with lyrics that mention "connection" or "a better tomorrow" may subliminally connect such good feelings to the product. Similarly, a car ad featuring lyrics like "born to run" or "ride with me" sustains the illusion of freedom and adventure, hence increasing the desirability of the product. A Nielsen research (2015) found that commercials with familiar music, particularly with lyrics, tended to evoke stronger emotional responses and worked better in viewer engagement compared to commercials without music. The study compared music ads and non-music ads across four categories: creativity, empathy, emotive power, and information power. As seen in Figure 2, music ads consistently perform better, especially in empathy and emotive power, demonstrating the strong emotional influence of music in advertising.

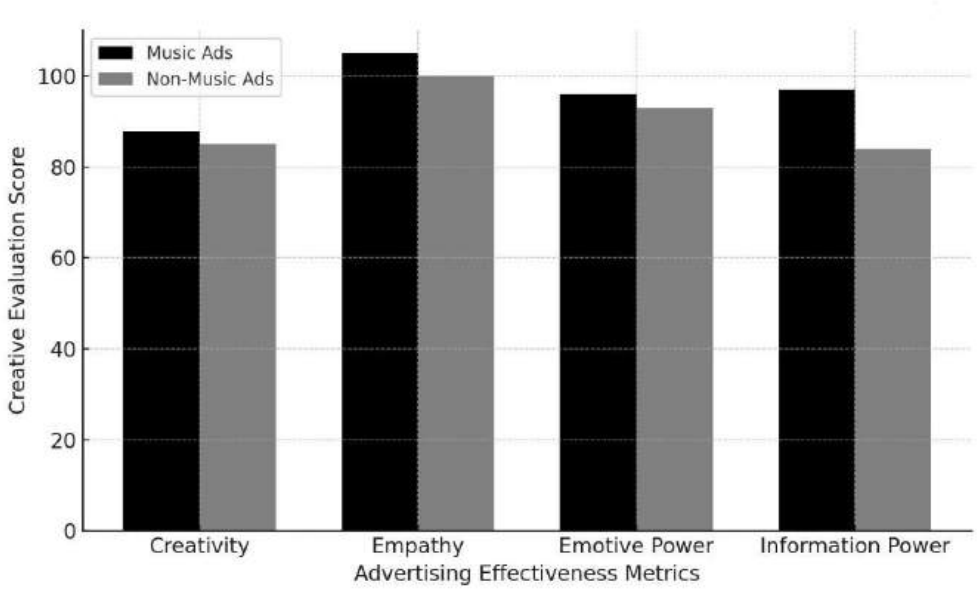


Figure 2. Impact of music in advertising on key effectiveness metrics

Song lyrics also have a vital function in film, television, and online media. Through their inclusion in soundtracks, background scores, and musical performances in films, lyrics enhance emotional resonance and consolidate character development. They help enrichen the story by bringing emotional context and depth. Lyrics also provoke sympathy, so audiences can more easily empathize with characters and what they're experiencing. They add richness to thematic content by cementing key messages and moods in a scene. A classic instance includes the use of Simon & Garfunkel's "The Sound of Silence" in The Graduate (1967), as it furthered the alienation and existential anguish of the principal character. Netflix's "13 Reasons Why" more recently fused emotionally charged tracks with lyrics echoing the psychological demons of its actors to raise consciousness regarding suicide and bullying (Brooks et al., 2018).

Beyond entertainment purposes, song lyrics have been employed in social movements and cultural expression for years. Billie Holiday's "Strange Fruit" was a haunting protest song against racial injustice, and Childish Gambino's "This Is America" used symbolic lyrics to condemn systemic racism and gun violence. Vuoskoski and Eerola (2012) found that listening to emotionally engaging lyrics can increase pro-social behavior and empathy. Their research highlights how music influences emotional responses and social connections. It suggests that media producers can use music to build awareness and enhance social sensitivity.

Moreover, advertising that incorporates music is particularly powerful among youth consumers. Teenagers and young adults are likely to be emotionally attached to songs, making them more receptive to promotions which involve personalized or familiar lyrics. Youth-oriented brands especially employ lyric-based advertising campaigns to target their consumer base. They can evoke nostalgia with well-known songs that remind consumers of happy times in the past. Some campaigns create hype with catchy, popular music that excites and engages listeners. Others use aspirational advertising with lyrics that inspire and urge consumers to strive for a particular

lifestyle. In addition, knowledge of the power of lyrics in media can lead to more critical consumption of the media they use. Whether discovering the way music enhances story in movies or the way protest songs reflect social concerns, students are able to gain valuable insights. They learn how lyrics assist in providing emotion and punch to stories in the media. More than that, they get to examine the way music is a voice for social and cultural movements, understanding on a deeper level what makes it such an effective force to shape society.

Personal and Academic Applications of Lyrics as Emotional Tools

Beyond the workplace, listeners, especially young ones, can transfer their understanding of song lyrics' emotional dynamics to their own lives. Teenagers in high school, for example, create playlists based on emotional need. While some playlists motivate, with uplifting lyrics to boost drive, some unwind, with calming music to reduce stress. Others play contemplative songs, using lyrics to sort through emotions and become more self-aware. For instance, a student preparing for a sports event might play a motivational anthem like “Eye of the Tiger”, whose lyrics emphasize resilience and perseverance (“Went the distance, now I’m back on my feet”). Conversely, a student feeling overwhelmed with schoolwork might turn to a song with reassuring lyrics, helping them process emotions and reduce stress.

As described in earlier sections, scientific research has consistently demonstrated that music is a part of the emotional equilibrium, supporting both physiological and psychological well-being. Research has also shown that whenever an individual is listening to music that he/she loves, either calming classical compositions or motivational pop songs, the bodies relax in response, showing decreased cortisol activity, thereby reducing stress and calmer emotional state (Chen, 2023). This is a valuable biological response that is particularly helpful for those dealing with ongoing difficulties controlling emotion as music provides a natural and easy mechanism for handling such feelings. Furthermore, research has established that music enhances attention and concentration and is therefore a helpful tool during learning or taking examinations. Appropriate music, normally instrumental or with soothing, repetitive phrases, can have the ability to bring its listeners into flow, improving their ability to hold information and perform tasks accurately.

Apart from its immediate physiological effect, music also influences self-concept and emotional resilience among young people. The majority of teenagers form intense emotional connections with specific songs and lyrics that reflect their own experiences and aspirations. Some adopt meaningful lyrics as personal mantras: short, memorable phrases that capture strength, inspiration, or determination. For example, a song with perseverance lyrics, i.e., "What doesn't kill you makes you stronger" (Kelly Clarkson), can serve as a comforting reminder to keep going through difficult times. In the same vein, music that tackles issues of self-acceptance and mental well-being, like "You are enough just as you are," can give the much-needed encouragement to individuals struggling with issues of identity. Such psychological support facilitates the development of emotional resilience, enabling individuals to overcome life's obstacles with more confidence. By integrating music into daily existence, humans can use lyrics

as a source of strength and inspiration. Therefore, music is more than entertainment; it is an effective emotional support system.

For those who enjoy reading and writing, song lyrics offer a new and accessible way of studying poetic devices in a modern context. Instead of merely reading classic poetry, students can discover rhyme, metaphor, and symbolism in the lyrics of contemporary songs. For those who enjoy psychology, knowing how music influences emotions and memory provides valuable insights into human behavior and mental processes. It is among what accounts for why certain songs trigger vivid memories or why music can change one's mood. For aspiring media practitioners, analyzing lyrics as cultural and commercial texts shows how music influences society, influences trends, and is strategically utilized in advertising and entertainment. In all these contexts, song lyrics present themselves as an interesting subject of study, the intersection of literature, psychology, and media in a meaningful way.

Lastly, song lyrics connect the general academic study and personal interest, making it an excellent subject of academic study as well as daily reflection. Reading and interpreting lyrics allow students to explore various disciplines, such as literature and history, in an engaging and meaningful way. Whether they are reading the symbolism in a traditional ballad, examining the emotional depth of a breakup song by a popular singer, or examining the social critique of a rap song, they are learning critical thinking and becoming more aware of art. Lyrics give us insight into trends in culture common themes. So they are a compelling tool for teaching that can be applied to any discipline. By analyzing song lyrics, students not only refine their critical thinking but also have a clearer sense of human emotions and society, reinforcing the conviction that music is much more than entertainment.

Conclusion

Lyrics are not just words in a song but an active, dynamic emotional and cultural practice. Observing how lyrics are used in emotion in various studies demonstrates the profound influence of lyrics. Whether used for self-reflection or social connection, lyrics are at the forefront of emotional experience. Consequently, music is no longer simply entertainment, but a means of learning, healing, and self-improvement.

The affective force of lyrics is in psychological and neuroscientific processes behind how we conceptualize and respond to music. They form autobiographical memories, reminding us of what happened before. They engage language and emotion brain networks, consolidating their impact. That is why lyrics get so personal and how they are utilized in therapy, advertising, and more, demonstrating their widespread use in human emotion and behavior.

Contextual and cultural factors also affect interpretation and experience of lyrics. Emotional responses are dependent on individual preferences, cultural background, and genre of music, but those which explore common issues like love, loss, or defiance can cut across languages, confirming their role in shaping shared experiences and social meaning. As research continues to explore the interplay of music, language, and emotion, it will continue to shed light

on how lyrics influence psychology, enhance cultural identity. Music and lyrics people of diverse backgrounds, providing a soundtrack to shared human experiences and emotions.

From Craft to Content: How Narrative Shapes Preference in Fashion Consumption
By Shanzay Rahman

Abstract

The paper brings in a socio-psychological perspective to investigate whether, under neoliberalism, Gen Z has developed a tendency to value narrative over craft. If so, then to what extent is this consumer psychology manufactured? This paper does so by drawing on Jean Baudrillard's theory of symbolic consumption and Pierre Bourdieu's discourse on cultural capital to investigate the psychological processes underlying fashion consumption practices, while also analysing how class and community identities shape the consumer's psychology. For primary sources, this paper analyses social media handles of Mango and Burberry to substantiate the marketing techniques that define and influence consumer psychology and argues that sign meaning and putting a value on that sign that determine the individual's sense of self in the social hierarchy lead towards a shift from 'content' to 'craft' in consumer preferences. Thus, brands like Mango and Burberry use this dual process in their marketing techniques to achieve an edge. From production to marketing vis a vis using cheap raw material to maximise surplus while at the same time using a blend of sign meaning and appeal to cultural capital to shift the focus from the contents of fashion to the narrative of fashion that helps the consumer to associate with the product on the level of self while also locating themselves at a certain social hierarchy, is the core of neoliberal marketing and production technique that this article substantiates by examples of campaigns and advertisements run by the brands like Mango and Burberry.

Keywords - Consumption Practices, Sign Value, Cultural Capital, Narrative Meaning, Gen Z.

Introduction

The rise of Industrialization and mechanical reproduction has greatly influenced consumer culture since the onset of modernization. Narrative meaning and association have greatly influenced consumerism, and purchasing has become an increasingly individualized act, replacing need-based purchases. The shift from a production process that relied on artisans and guilds to assembly lines had serious implications for the quality and aesthetics of the end product. According to cultural theorist Walter Benjamin (1936), the process of 'mechanical reproduction' involves a certain 'Aura decay.' Thus, with modernization, we see a development of the advertising industry to compensate for the aura decay and superimpose a meaning, thus making it synonymous with the product**.** As this paper would go on to show, that superimposed meaning or narrative is essential in the modernist development to make way for commercialized assembly line production. Since the onset of this process, it has repeatedly evolved only towards cheaper, unsubstantial production, and, to compensate for it, the capitalist market has developed a mammoth marketing machinery to superimpose meaning onto degrading products.

Fashion, Consumption and Capitalism

Capitalist consumption patterns also influence the rise of the modern fashion industry. Since modernity, fashion has increasingly blended craftsmanship and storytelling into its definition, thereby setting cultural trends that appeal to individuals' sense of belonging and their position within society. However, since globalization and the rise of neoliberalism, manufacturing has moved increasingly away from craft and towards narrative. The fashion industry is no exception to that; fashion has moved towards cheap production to minimise expenditure (Jin & Shin, 2021), while spending much more on marketing to shroud the product in a narrative that creates a certain 'association appeal' for the individual. In other words, a fashion product conveys a person's status, culture, and belonging. Now, with the dawn of Neoliberalism, the world has witnessed the rise of Gen Z as almost a new cultural phenomenon that articulates itself in the young bodies and minds. This paper seeks to analyse how Gen Z and its sensibilities have been shaped by marketing and cultural trends disseminated through social media, thereby establishing a preference for content over craft. It does so by analysing marketing trends and narrative patterns on social media, particularly in the short-form content of select fast-fashion brands, namely Mango and Burberry. According to theorists, Gen Z is a universal phenomenon, sharing common values and bounded by a specific 'technologies of self' (De Witte, 2022). However, central to this paper's inquiry is also understanding how far Gen Z's sensibility is organic.

The paper brings in a socio-psychological perspective to investigate whether, under neoliberalism, Gen Z has developed a tendency to value narrative over craft. If so, then to what extent is this consumer psychology manufactured? This paper does so by drawing on Jean Baudrillard's (1998) theory of symbolic consumption and Pierre Bourdieu's (1986) discourse on cultural capital to investigate the psychological processes underlying fashion consumption practices, while also analysing how class and community identities shape the consumer's psychology. For primary sources, this paper analyses social media handles of Mango and Burberry to substantiate the marketing techniques that define and influence consumer psychology.

Fashion, Modernity, and Symbolic Consumption

In Pre-capitalist societies, social values attached to goods were significantly less, production was need-based, and only happened according to the demand of the market (Polanyi, 1944). The concept of market creation is fairly absent from pre-capitalist societies; usually, guilds would receive a commission for production that they would deliver, unlike modern practices, where production happens on market prediction. Thus, such a production practice was one-dimensional as it operated on a need-to-produce basis, serving utility. Consumption practices were also need-based, and products were perceived on their material meaning with no symbolic imposition like in modern consumption practices. As society moved through capitalism, social, cultural, and symbolic meaning got attached to products largely as a result of a vast operational system of marketing and advertising (Baudrillard, 1998). To demonstrate the point, in a pre-capitalist society, a horse-drawn carriage was intended for transportation and moving of

goods, while today a modern sports bike, which may also be used for transportation, carries with it a plethora of symbolic connotations. A modern sports bike and its consumption rationale are built not around transportation but a sense of masculinity, a larger-than-life experience owing to its fast pace, and the fact that other people would consider the rider as '*cool*'.

Jean Baudillard has argued that in modern societies, materials are often conceived in terms of their sign value rather than their utility. He further stressed that it is through signs and meanings of goods that people in a given society convey and find their status, belongingness, and identity (Baudrillard, 1998). Thus, in the advanced capitalist mode of production, objects often surpass their intended utility to become markers of social and cultural meanings that help an individual shape their identity. Arguably, Gen Z, whose perception is largely based on social media exposure, thus builds their perception and self around object association that grounds their identity in a consumerism-driven world of alienated beings. An iPhone brings about status vs. an Android phone; having a certain limited edition jersey made out of cheap polyester brings about a sense of uniqueness to the wearer. Pierre Bourdieu's theory of cultural capital further substantiates how objects with symbolic values produce a certain sense of hierarchy in terms of cultural capital that sets the owner of a certain object higher within the social hierarchy than the one who lacks the object (Bourdieu, 1986). Fashion Brands like Mango, in their virtual presence, target their campaign in promoting a certain sense of worth, self, and culture that can only be achieved by possessing their products. Fashion serves as an important mechanism through which cultural capital is achieved and displayed. An individual's clothes communicate their education, social status, and aesthetic sensibilities.

Campaigns and Gen Z consumption

Since the COVID years, the rise in short-form content (Blackford & Hamilton, 2025), has increasingly achieved a hegemonic cultural shift in creating a global culture of consumption. Arguably, this consumption culture is heavily dependent on symbolic and narrative meaning that appeals to individuals by inflicting a sense of cultural association, which helps the individual to locate herself in a specific social hierarchy. Such a way of looking at consumption culture and the narrative meaning of products locates itself at the intersection of Baudrillard's (1998) theory of Sign Value and Bourdieu's (1986) theory of cultural capital. Both these processes, sign meaning and putting a value on that sign, determine the individual's sense of self in the social hierarchy, leading towards a shift from 'content' to 'craft' in consumer preferences. Thereby, Brands like Mango and Burberry use this dual process in their marketing techniques to achieve an edge. From production to marketing vis a vis using cheap raw material to maximise surplus while at the same time using a blend of sign meaning and appeal to cultural capital to shift the focus from the contents of fashion to the narrative of fashion that helps the consumer to associate with the product on the level of self while also locating themselves at a certain social hierarchy, is the core of neoliberal marketing and production technique that this article substantiates by examples of campaigns and advertisements run by the brands like Mango and burberry.

- **Burberry**

Burberry's 2025 campaign "It's always Burberry weather" had a massive cultural appeal of situating the trench coat in the mundane, everyday interaction (Burberry, 2026). Commenting on London's rain, the gloomy narrative of a rainy day was transformed into a subtle celebration of Burberry trench coats. The campaign transformed the pure elite ties of the trench coat into everyday Londoners' experience by situating it in places like narrow alleys and convenience stores while branding it through bodies of recognized actors such as Kate Winslet. By locating the trench coat in everyday British experience, the symbolic value was placed on authentic Britishness, celebration of rain through compelling storytelling as opposed to a narrative of raw wealth. While employing recognized actors to justify its cultural value to the prospective purchaser. To further the marketing deep into consumers' psyches, the advertisement compares the trench coat as a faithful companion who doesn't let go even when things get gloomy. The video starts by twisting Shakespeare's ' Shall I compare thee to a summer's day'(Shakespeare, 2017) on its head and comparing love to rain and gloom. The narrative unfolds that it might be easier to love on a sunny day, but when it comes to a rainy day, when one is down, then love becomes the sun. Thus, the campaign, on one hand, plays on love to bind the consumer with the product on a personal level while projecting the trench coat in everyday mundane reality juxtaposed against the cultural heritage of London; moreover, it showcases the trench coats on bodies of celebrities to derive future cultural appeal and credibility, an intricate play at consumer psychology.

- **Mango**

The 2025 Supriya Lele X Mango exclusive collaboration portrayed a hyperbold, unapologetic feminist expression in a dystopian setting juxtaposed against extreme solitude and an estranged self - "A collection that manages to convey contemporaneity while showcasing Supriya Lele's bold vision of femininity: dresses with character, suggestive transparencies, sensual details and a vibrant colour palette that invites you to express yourself unapologetically" (Mango, 2026). The video hinges on Gen Z's compassion for estrangement, solitude, alienation and a bold, unapologetic self. By locating the performer in a dystopian setting, the video messages to the consumer, while the use of flashing lights to highlight the transparencies of the body transforms the clothes into statements of autonomy and resistance. The clothes appear to be just complementing the body, not hiding it, empowering, not defensive, liberating, not restrictive. On the other hand, a collection by an avant-garde artist cum designer like Supriya Lele gives the wearer the sense of a collector; the fashion item is not just a piece of cloth; it transcends into a piece of art that clings to the body and assigns meaning, capital, taste, and distinctiveness. The wearer transforms herself into the artist's imagery as her body becomes the canvas on which the artist has painted. Thus, the narrative situates the fashion articles in the sign value of reclaiming one's own body and autonomy while also bringing in the cultural legitimization of avant-garde art through which the wearer can hold herself in a specific niche of subjectivity and a social hierarchy.

Conclusion

This research analysis has investigated the consumer psychology of Gen Z and the marketing techniques employed by brands like Mango and Burberry to show how, in the neoliberal setting, marketing tends to present a narrative that situates the product within a specific sign and cultural value. Gen Z's preference for *Craft* over *Content* is not an organic development. Still, it must be traced within the historical development of production processes and within a sociological perspective that analyses the sign and cultural value of a purchase. The article revealed that Mango, in its collaboration campaign with Supriya Lele, utilized a bold feminist narrative in a dystopian setting to associate with a specific Gen Z mentality while utilizing the cultural appeal of avant-garde art to derive its cultural legitimacy. The dual process of sign value and cultural value in marketing narrative is evident in Burberry's 'It's always Burberry weather' campaign too. Burberry utilized the sign value of love, rain, and London heritage through compelling storytelling and in a mundane portrayal of everyday life while deriving its cultural legitimacy from Shakespeare and by starring celebrities. Thus, the modern fashion industry operates by inculcating the preference for narrative over craft and, in that process, transforms the product from its mere utilitarian value to a living, breathing cultural artifact of layered meaning and socio-cultural ties.

Works Cited

Baudrillard, J. (1998). *The Consumer Society: Myths and Structures*. Sage.

Benjamin, W. (1936). The Work of Art in the Age of Mechanical Reproduction. *Journal of Philosophy of Education*, *41*(3), 363–380. https://doi.org/10.1111/j.1467-9752.2007.00579.x

Blackford, C., & Hamilton, M. (2025). *No Turning Back? Examining the Rise of Short-Form Content in Post-COVID Marketing for Gen-Z and Gen Alpha*. Digital Commons at Loyola Marymount University and Loyola Law School. https://digitalcommons.lmu.edu/honors-thesis/590/#:~:text=Coupled%20with%20a%20rise%20of,spans%20getting%20shorter%20and%20shorter.

Bourdieu, P. (1986). The Forms of Capital. *Journal of Economic Sociology*, *3*(5), 241–258. https://web.stanford.edu/~eckert/PDF/Bourdieu1986.pdf

Burberry. (2026). Youtube.Com. https://www.youtube.com/watch?v=zvzqBEDtlDo&list=PLY7E_CZZ_9_O1xVfU3O58ekBSIH1L-5bz

De Witte, M. (2022, January 3). *What to know about Gen Z*. News.Stanford.Edu. https://news.stanford.edu/stories/2022/01/know-gen-z

Jin, B. E., & Shin, D. C. (2021). The Power of 4th Industrial Revolution in the Fashion industry: What, why, and How Has the Industry changed? *Fashion and Textiles*, *8*(1).

Mango. (2026). *Supriya Lele x Mango* [Review of *Supriya Lele x Mango*]. Youtube.Com. https://www.youtube.com/watch?v=7tFEEph99JY

Polanyi, K. (1944). *The Great Transformation*. Beacon Press. https://www.goodreads.com/en/book/show/53981.The_Great_Transformation

Shakespeare, W. (2017, April 25). *Sonnet 18: Shall I compare thee to a summer's day?* [Review of *Sonnet 18: Shall I compare thee to a summer's day?*]. The Poetry Foundation. https://www.poetryfoundation.org/poems/45087/sonnet-18-shall-i-compare-thee-to-a-summers-day

Comparing Exoplanet Transit Detection Pipelines Under Realistic Noise and Data Gaps By Kyle He

Abstract

Detecting exoplanet transits in brightness time series is complicated by time-correlated (red) noise and missing data, which can hide real transits and bias recovered parameters. The choice of how to remove stellar trends before a periodic transit search is therefore consequential, yet the relative merits of common detrending strategies under controlled, realistic conditions are not obvious. We built an injection-recovery framework that injects known limb-darkened transits into 30 real TESS light curves (real photometric baselines, not simulated), then adds controlled white noise, correlated noise, and synthetic data gaps whose amplitudes and severities we set as experimental variables, across a 2×2 grid of noise and gap severity, and benchmarks five analysis pipelines that share a common periodic-search framework but differ in how trends and transit signals are modeled: a rolling-median baseline (P0), ridge feature-regression (P1), masked long-trend with optional wavelet denoising (P2), a Gaussian-process detrend (P3), and a joint trend-plus-transit model (P4). Recovery was graded on period, mid-transit time, and depth within fixed tolerances on an identical set of 2400 trials, and pipelines were compared with paired (McNemar) exact tests. Of the four alternatives to the baseline, the joint model (P4) and the Gaussian-process detrend (P3) improved overall recovery, the wavelet pipeline (P2) was essentially neutral, and the feature-regression pipeline (P1) was significantly worse. P4's advantage was broad across individual stars and concentrated in the high-noise regime, with its largest gains in mid-transit timing; P3's pooled improvement was driven by a minority of stars. A complementary null-trial analysis showed that, under a common depth-based detection statistic, P4's recovery advantage did not extend to distinguishing transits from noise, illustrating that recovery accuracy and false-positive resistance are distinct axes of pipeline performance.

1. Introduction

The transit method detects an exoplanet by the small, periodic dip in a star's brightness as the planet crosses in front of it. In practice the dips are shallow, often a fraction of a percent, and the brightness time series in which they are sought contains far more than random measurement scatter. Real light curves carry slow instrumental drifts, time-correlated stellar variability known as red noise, and gaps where data are missing. These effects can bury a genuine transit, distort the recovered period, timing, or depth, and even produce convincing spurious signals. Mitigating them is a central problem in transit photometry (Pont et al. 231).

Most transit searches first remove the stellar and instrumental trend (“detrending”) and then run a periodic search such as Box Least Squares (Kovács et al. 369) for a repeating dip. The detrending step is where much of the difficulty lives: remove too little and the trend masquerades as or hides a transit; remove too much and the transit itself is flattened away. Many detrending strategies exist, from simple running medians to Gaussian-process models of correlated noise (Foreman-Mackey et al. 220) to joint models that fit the trend and the transit simultaneously.

Which strategy is most reliable, and under what observing conditions, is not obvious, and the answer plausibly depends on how severe the noise and gaps are.

This work addresses that question with a controlled injection-recovery experiment. We inject transits of known period, mid-transit time, duration, and depth into real TESS (Barclay et al.) light curves, corrupt them with controlled white and correlated noise and with data gaps at two severity levels each, and then ask how reliably each of five detection pipelines recovers the injected parameters. Correlated, stellar-variability-driven noise on transit-like timescales is a recognized limit on transit photometry (Morris et al.), and is the regime our high-noise condition is designed to probe.

Injection-recovery testing, in which fake transits of known properties are added to real light curves and a pipeline is asked to find them, is the standard way to test a transit pipeline. Almost always, though, it has been used to measure how well a single pipeline performs, not to compare several against each other. The Kepler mission set the pattern: the Christiansen et al. series (Christiansen et al. art. 35, art. 95, art. 99) injected signals into Kepler data to measure how the chance of detection grows with signal strength, a number needed to estimate how common planets are, and other surveys have done the same. In every case the pipeline stays fixed and the injected signal is varied, so the question is "how deep must a planet be before we can find it?", not "which analysis method finds it best?".

Two earlier studies come closer to what we do. Hippke et al. ("Wōtan") compared many different detrending methods on Kepler, K2, and TESS data and found that one particular smoothing filter usually works best, a recommendation now in wide use. Like us, they held the search step fixed and varied only the detrending; unlike us, they mainly asked whether the signal was recovered, and did not test whether the differences between methods were statistically significant, break the results down star by star, or check how often each method reports a transit when none is present. Separately, several groups have asked whether fitting the trend and transit together beats removing the trend first and searching afterward. After Foreman-Mackey et al. ("A Systematic Search for Transiting Planets") set out how to fit both at once, Kovács et al. ("Optimal Trend Filtering") found no benefit on K2 data, and Taaki et al. found only a small, non-significant improvement on TESS data. So whether joint fitting helps is still open, with the evidence so far pointing mostly to "no clear difference".

What is new here is not injection-recovery testing itself, nor the observation that detrending matters, but the comparison design. Every pipeline sees the same 2400 trials, so we test the differences between them with matched (paired) statistics rather than comparing separately drawn samples; we vary noise and gap severity in a 2×2 grid, so an advantage can be traced to the conditions that produce it; we grade period, timing, and depth separately, so a failure can be pinned to a specific parameter; we check each pooled result star by star, which separates a broad advantage from one driven by a few light curves; and we run a matched set of injection-free trials, which distinguishes how well a pipeline recovers a real transit from how well it avoids "finding" one that is absent. The star-by-star and no-transit analyses are what mainly set this work apart from Hippke et al. ("Wōtan"), and the paired design across noise levels is what sets it apart from the joint-fitting studies above.

We went in expecting the joint model (P4) to recover transits best in the hardest conditions, and the wavelet (P2) and Gaussian-process (P3) detrends to improve on the simple baseline (P0) where drifts dominate. As reported below, the results partly confirm and partly complicate these expectations, and a complementary analysis of injection-free "null" trials shows that the pipeline which recovers transits best is not necessarily the one that best tells them apart from noise.

2. Methods

2.1. Overview and Workflow

The experiment has five stages, each reproducible and auditable. We select a sample of stars and download their real TESS light curves; for each star we generate many trials by normalizing the light curve, injecting a limb-darkened transit with known parameters, and adding controlled synthetic noise and data gaps, recording each trial's injected "truth" values and condition labels. We then run all five pipelines on the identical set of trials, grade each recovered result against the injected truth with fixed tolerances, and compare the pipelines with paired statistics, supplemented by a per-star robustness check and a null-trial false-positive analysis. Figure 1 shows a representative recovered trial.

The central design choice is that all five pipelines are evaluated on the same trials. This makes the comparison paired: for any two pipelines and any trial, both either recovered it or did not, so differences in their success counts reflect the pipeline design rather than which trials each happened to see. Throughout, we cross two noise levels (low, high) with two gap levels (minimal, severe) to form a 2×2 grid of observing conditions, and report results both pooled and per cell.

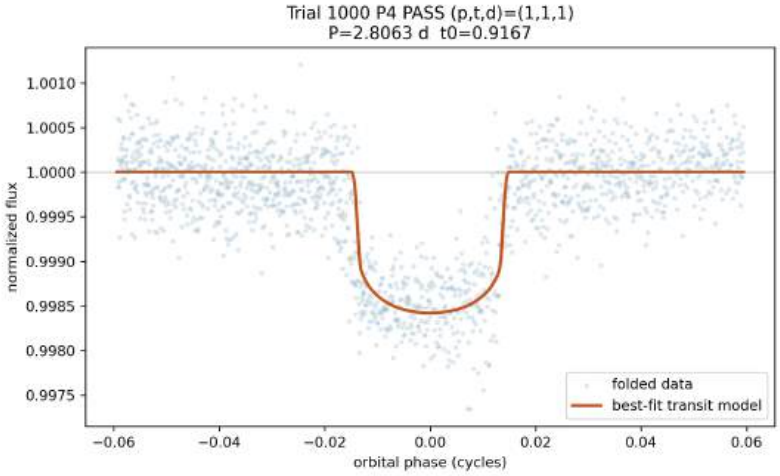


Fig 1: Example phase-folded trial after recovery (one representative trial). Light points are the folded data on the recovered period; the solid curve is the recovered limb-darkened transit model at the fitted depth.

2.2. Stellar Sample and Data

We selected 30 stars from the TESS Objects of Interest catalog by querying the NASA Exoplanet Archive, restricting to targets with confirmed or known planet dispositions and a TESS magnitude brighter than 12.5. We sorted the qualifying targets by brightness, removed duplicates so no star appears twice, and kept the 30 brightest. The selection is therefore fixed rather than random: it involves no hand-picking and can be reproduced exactly by rerunning the same query, but it is a deliberate brightest-first cut, not a sample meant to represent the full range of TOI targets. Bright, confirmed-planet hosts were chosen because their light curves are well-characterized and high in quality, so that the baseline data into which we inject transits are representative of real survey targets rather than pathological cases. This choice has a consequence worth stating. Because these are some of the best light curves TESS produces, the recovery rates we report are probably better than the same pipelines would achieve on a typical, fainter set of targets, and the ranking is established only in this favorable setting. Fainter, more variable, or more crowded stars make trend removal harder, and we cannot claim from these data that the same ordering would hold there. For each star we retrieved a TESS light curve through the MAST archive (Mikulski Archive for Space Telescopes) using Lightkurve (Lightkurve Collaboration et al.), requiring Science Processing Operations Center (SPOC) products at 120 s cadence; all 30 selected stars had such products (spanning TESS sectors 1 through 98), so the sample is uniform in pipeline provenance and cadence. SPOC products provide calibrated, systematics-corrected photometry at uniform cadence.

From each light curve we kept the time stamps and the PDCSAP (Pre-search Data Conditioning Simple Aperture Photometry) flux, which has had common-mode instrumental systematics removed by the SPOC Presearch Data Conditioning module. We used PDCSAP rather than simple aperture photometry (SAP) deliberately: SAP retains large, star-specific instrumental signals (pointing drift, scattered light, thermal ramps) that vary substantially from target to target, so injecting into SAP would let an uncontrolled real systematic, differing across stars, dominate each trial and tangle pipeline differences up with which star had the worst leftover signals. PDCSAP provides a comparatively clean and uniform baseline onto which the controlled white, correlated, and gap perturbations of Section 2.4 can be imposed as the experimental variable. The consequence is that the pipelines here compete only on what remains, mainly leftover stellar variability plus the correlated noise we add; we defer the implications of this for interpreting the results to the Limitations (Section 4.6).

We discarded non-finite samples, sorted by time, divided the flux by its median so the out-of-transit level is near unity, and applied an 8σ robust clip using the median absolute deviation (MAD) to remove gross outliers while preserving genuine transit depths (which are far shallower than 8σ of the cleaned scatter). Using real time stamps preserves each star's true cadence and its native observational gaps, so that injected transits are sampled exactly as a real transit on that star would be.

2.3. Transit Injection Model

Rather than a rectangular (“box”) dip, we injected limb-darkened transits so that the recovered transit shape would resemble a real planetary transit, whose floor is rounded because

the stellar disk is brighter at its center than its limb (Fig. 2). We adopted a quadratic limb-darkening law for the specific intensity as a function of the cosine of the angle from the disk center, μ,

$$I(\mu) = 1 - u_1(1-\mu) - u_2(1-\mu)^2 \qquad (1)$$

with coefficients $u_1 = 0.35$ and $u_2 = 0.20$, representative values for a Sun-like star (Kipping 2152), and impact parameter b = 0.30 (the projected sky distance between the planet and stellar centers at mid-transit, in stellar radii). We used the same coefficients for every star, a fixed modeling choice rather than a per-star fit. Using one fixed transit shape means every pipeline faces the identical target on every trial, so differences in recovery cannot come from the shape changing between stars. The trade-off is that this defines a deliberately narrow benchmark: every trial is drawn from a single, closely specified transit family, with the shared depth template below matched to the injected shape by design, so the benchmark does not probe how the pipelines handle transits whose shape departs from the injected one, a limitation we return to in

Section 4.6.

To compute the flux loss as the planet crosses the disk, we numerically integrated the obscured intensity. For a planet-to-star radius ratio k and a center-to-center sky separation d (both in stellar radii), the fraction of the total stellar flux blocked is

$$\delta(d,k) = \frac{1}{F_\star}\int_0^1 2\theta(r;d,k)\, I(r)\, r\, dr, \qquad F_\star = \pi\left(1 - \frac{u_1}{3} - \frac{u_2}{6}\right) \qquad (2)$$

where I(r) is the intensity at projected radius r from Eq. 1 with $\mu = \sqrt{(1-r^2)}$, and 2θ(r;d,k) is the angular extent of the annulus at radius r that lies under the planet's disk (zero outside the overlap, 2π where the planet fully covers the annulus, and an arccosine of the standard two-circle overlap geometry in between). We precomputed δ(d,k) on a grid of k and d once per run (the grid resolution is a fixed configuration parameter) and interpolated within it, which is far faster than integrating per time stamp. For each trial the time-dependent sky separation was mapped from the orbital phase assuming a central chord scaled by the impact parameter, and the resulting δ(t) was rescaled so that the flux drop at mid-transit equaled the trial's target depth. The transit was then applied multiplicatively to the normalized baseline, $f(t) \rightarrow f(t)[1 - \delta(t)]$.

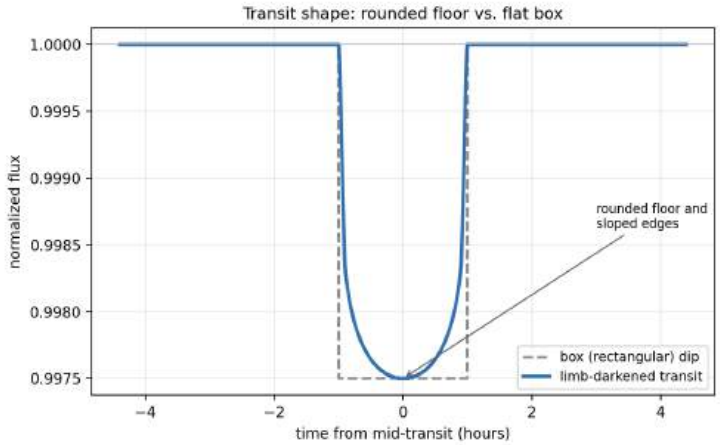


Fig 2: Why we inject a limb-darkened transit rather than a flat "box" dip. Because a real star is brighter at its center than at its edge, a planet crossing it produces a dip with a rounded floor and sloped sides (solid), not the flat bottom and vertical walls of a box (dashed). Both have the same depth and duration here. Injecting the more realistic shape makes the recovered transits look like real ones.

For each trial the period P was drawn uniformly from 2 to 8 days, the mid-transit depth uniformly from 0.0008 to 0.0030 in fractional flux (shallow, comparable to small planets), and the total transit duration fixed at 2 hours. The mid-transit time t_0 was drawn uniformly within the window but rejected and redrawn (up to 200 retries) unless at least two transit events fell within the observing baseline, which prevents trials that are unrecoverable in principle (no sampled transit) from inflating the failure rate of every pipeline equally and washing out real differences. We adopted the limb-darkened injection in place of a simpler box-shaped dip, and a common template-based depth estimator in place of a median difference, because the limb-darkened form is more physically realistic and, as described below, lets every pipeline estimate depth against the same template.

2.4. Noise and Gap Models

Two noise components were added to each injected light curve; Fig. 3 shows the full construction of one trial stage by stage. White (uncorrelated) noise was drawn from a zero-mean Gaussian with standard deviation set by the noise level. Time-correlated ("red") noise was generated as an Ornstein-Uhlenbeck (OU) process, the continuous-time analogue of a first-order autoregressive process, which is an appropriate formulation on the irregular TESS time sampling. The OU process satisfies, between consecutive samples separated by Δt,

$$x_i = \alpha\, x_{i-1} + \sqrt{1-\alpha^2}\,\sigma_{\mathrm{red}}\,\varepsilon_i, \qquad \alpha = e^{-\Delta t/\tau} \tag{3}$$

where τ is the correlation timescale (0.08 d, about 1.9 hours, comparable to the transit duration), ε_i is standard normal, and the $\sqrt{(1-\alpha^2)}$ factor keeps the process stationary with standard deviation σ_{red} regardless of the local cadence. We deliberately set τ near the transit duration because correlated noise on that timescale is the hardest to deal with: it looks like the very feature we are searching for, and it is exactly the case where a model of correlated noise should help if it helps at all. The two noise levels differ in their amplitudes: the white standard deviation is 2.5×10^{-4} (low) or 7×10^{-4} (high) in fractional flux, and the correlated amplitude σ_{red} is 2.0×10^{-4} (low) or 7×10^{-4} (high). For reference, the injected depths (0.0008–0.0030) span roughly one to twelve times these per-point noise amplitudes, so recovery is non-trivial in the high-noise regime.

On top of each light curve's native gaps we imposed additional missing-data blocks: the minimal level removed 2 contiguous blocks of 0.2 d each, the severe level 6 blocks of 0.6 d each, with block start times drawn uniformly. Compared with the average length of data we have per star (27.1 d, about one TESS sector), this removes roughly 1.5% of the data at the minimal level and 13.3% at the severe level. For context, the light curves already have natural gaps covering about 16% of their length, mostly the regular pause for data downlink each orbit plus losses to

scattered light. So the minimal setting is gentler than the interruption a real sector already has, while the severe setting roughly doubles the total missing fraction; the two levels are meant to bracket realistic conditions rather than reproduce one exactly. Block removal (rather than random point dropping) was used because real gaps are contiguous, arising from data downlinks and spacecraft operations, and contiguous gaps are harder for detrending and periodic search than the same number of scattered missing points. This also means gap structure, not just the missing fraction, shapes the difficulty: the same fraction removed as many small scattered gaps would stress the pipelines differently, and because we hold the structure fixed at realistic contiguous blocks, our gap-severity axis varies the missing fraction at one representative structure rather than sweeping structure itself. Trials left with fewer than 300 surviving points were regenerated.

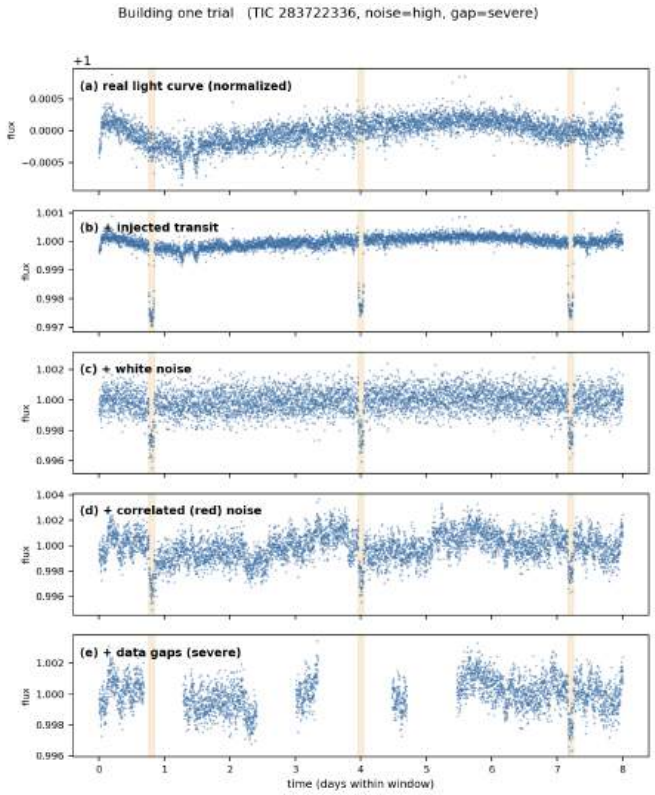


Fig 3: How one trial is built, shown on a real light curve. Starting from the normalized data (a), we add a known transit (b), then uncorrelated "white" noise (c), then correlated "red" noise (d), and finally remove blocks of data to imitate gaps (e). The shaded bands mark where the injected transits fall. This example uses the hardest settings (high noise, severe gaps); by the last panel the transit is difficult to see by eye, which is the regime the search has to cope with.

2.5. The Five Pipelines

All five pipelines accept the same (time, flux) input and return a recovered period, mid-transit time, depth, and a detection statistic under a shared schema. They share an identical periodic-search core and final depth estimator (described after the pipelines). The first four (P0–P3) differ only in the detrending step that precedes the search, isolating the effect of detrending strategy; the fifth (P4) replaces the detrend-then-search sequence with a joint

trend-plus-transit fit, so the benchmark also contrasts a sequential architecture with a joint one. Figure 4 shows the shared structure and where the pipelines differ.

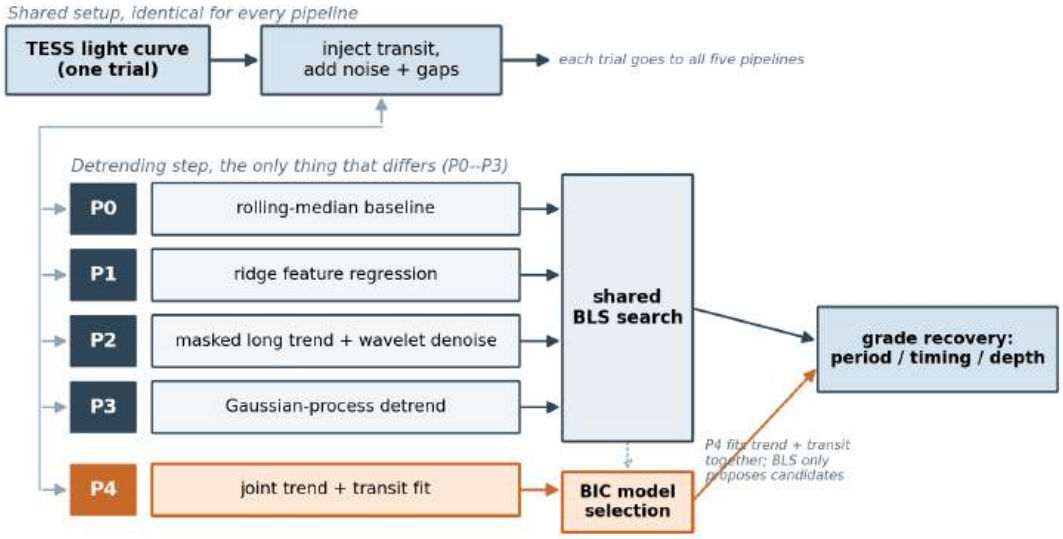


Fig 4: The five pipelines share the same setup, search, and grading, and differ only in how they remove the stellar trend (and, for P4, in fitting the trend and transit together). Each trial is built once and handed to all five pipelines. P0–P3 remove a trend and then run the same Box Least Squares (BLS) search; P4 instead uses BLS only to propose candidate signals and then fits the trend and transit jointly, choosing among candidates with the Bayesian Information Criterion (BIC). Every pipeline is graded the same way, on period, timing, and depth.

Shared periodic search. After detrending, each pipeline searches for a periodic transit with the Box Least Squares (BLS) algorithm (Kovács et al. 369), as implemented in astropy (Astropy Collaboration et al.), over the 2–8 day range at durations of 1, 2, 3, and 4 hours. (More elaborate searches such as Transit Least Squares (Hippke and Heller) use a fixed transit-shaped template; we use BLS as the shared search so pipeline differences come from detrending, not the search.) We use the same BLS search for all pipelines so that differences among P0–P3 come from detrending alone, and so that P4's joint fit is compared against a shared baseline. To guard against two common BLS failure modes, the search adds a harmonic check (re-evaluating at half and double the best period, since BLS often locks onto an integer multiple of the true one) and a local refinement (a fine period grid within ±2% of the best period). The recovered t_0 is the phase of the best configuration.

P0: rolling-median baseline. The flux is detrended by subtracting a centered rolling-median trend computed over a window of ~1.5 days, then passed to BLS. The rolling median is robust to outliers and to the transit itself (a brief dip barely moves a day-scale median), making this a simple, hard-to-beat baseline. It removes slow drifts but makes no attempt to model correlated noise.

P1: ridge feature-regression detrend. A quick BLS provides a provisional ephemeris; the in-transit points are masked, and a linear trend is fit to the remaining data by ridge regression.

The design matrix X holds a constant, low-order polynomials in time, sine and cosine terms at 5- and 10-day periods, and one indicator column per gap-separated segment (to absorb step-like offsets between observing blocks). The coefficients minimize

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|^2 + \lambda \sum_{j \in \mathcal{R}} \beta_j^2 \quad (4)$$

where the penalty applies only to the non-offset columns R (the constant and segment indicators are left unpenalized) and the design columns are standardized before solving so that λ acts comparably across features. The fitted trend is subtracted and BLS is run on the residual. The intent is that an explicit basis can capture structured variability a median misses, in the spirit of regression-based decorrelation methods (Luger et al.). A known risk of a flexible basis fitted to masked data is that residual transit signal can be absorbed into the trend; whether this occurs in practice is examined in the Results.

P2: masked long trend with optional wavelet denoising. Transits are masked using a provisional ephemeris, a long-timescale trend is built by interpolating across the masked points and applying a wide rolling median (so the trend cannot follow a transit), and the residual is optionally passed through a soft-threshold wavelet denoiser. Wavelet denoising decomposes the residual into multiscale detail coefficients and shrinks those below a noise-scaled threshold $\sigma\sqrt{(2 \ln N)}$ (the universal threshold, with σ the robustly estimated noise and N the series length) before reconstruction, suppressing incoherent high-frequency scatter while preserving coherent structure. BLS is then run on the cleaned residual.

P3: Gaussian-process detrend. Transits are masked, and the out-of-transit flux is modeled as a Gaussian process (GP) with a Matérn-3/2 covariance kernel,

$$k(\tau) = \sigma_{\mathrm{GP}}^2 \left(1 + \frac{\sqrt{3}\,\tau}{\rho}\right) \exp\left(-\frac{\sqrt{3}\,\tau}{\rho}\right) \quad (5)$$

where τ is the time separation, σ_{GP} the amplitude, and ρ the correlation length. The hyperparameters are fit by maximizing the GP marginal likelihood using the celerite2 library (Foreman-Mackey et al. 220), which evaluates the likelihood in time linear in the number of points. Crucially, ρ is constrained to lie between 0.7 and 30 d. The lower bound keeps the GP from absorbing the transit: the injected transits last 2 hours (0.083 d), so a floor of 0.7 d keeps the correlation length more than eight times longer than a transit, and a Matérn-3/2 process with ρ that large cannot follow the dip. Without it, the fit could lower its error by letting the GP trace the transit itself, the same concern that motivates the wide rolling median in P2. The maximum-likelihood GP trend is then subtracted and BLS run on the residual.

P4: joint trend-plus-transit model. Rather than detrend then search, P4 uses BLS only to propose a small set of candidate periods and epochs, then for each candidate jointly fits the trend and the transit. For a candidate (P, t_0) it builds the same trend basis X as P1 and appends one column equal to the negative limb-darkened transit template at that ephemeris, so the combined linear model is

$$y \approx X\beta + d\, s(t; P, t_0) \quad (6)$$

where s is the unit transit shape and the fitted coefficient d is the depth. This is solved by the same ridge procedure as Eq. 4. Each candidate is scored by the Bayesian Information Criterion,

$$\mathrm{BIC} = N \ln\left(\frac{\mathrm{RSS}}{N}\right) + m \ln N \quad (7)$$

with RSS the residual sum of squares, N the number of points, and m the number of fitted parameters; the candidate with the lowest BIC is reported. The BIC penalty discourages the flexible trend basis from spuriously improving the fit, balancing fit quality against model complexity. Fitting the trend and transit together, rather than sequentially, is what allows P4 to avoid the self-inflicted signal loss that a detrend-then-search pipeline can suffer when the trend model partly absorbs the transit. Because P4 differs from P0–P3 in both its detrending basis and this joint, BIC-selected fit, any performance difference it shows reflects the joint architecture as a whole rather than a detrending step in isolation.

Shared depth estimate. For all five pipelines, depth is finally estimated by a linear least-squares fit of the same limb-darkened template at the recovered ephemeris (baseline plus a scaled template), so that depth grading is measured the same way for every pipeline and is not biased toward whichever detrending strategy happens to estimate amplitude differently. The pipelines are also given the true transit duration and impact parameter, a deliberate simplification: a fully blind search would have to recover the duration too, which enlarges the search and adds failure modes (a wrong duration spoils both the transit mask and the depth template). Supplying these two numbers makes the benchmark a test of recovering period, timing, and depth given a known transit shape, isolating the trend model, so the recovery rates here should be read as best-case relative to a fully blind search.

2.6. Recovery Criteria

Each recovered result was graded against the injected truth with three checks, all fixed before the pipelines were run, so the grading could not be tuned to favor any pipeline:

- Period (p_ok): $|P_{found} - P_{true}|/P_{true} \leq 0.01$.
- Timing (t_ok): the period-wrapped difference between recovered and true mid-transit time is at most 1 hour (half the transit duration). Wrapping modulo the period treats solutions offset by an integer number of periods, which are phase-consistent and physically equivalent, as correct.
- Depth (d_ok): $0.70\, d_{true} \leq d_{found} \leq 1.30\, d_{true}$, a symmetric multiplicative window that tolerates modest depth bias from detrending while rejecting clearly wrong depths.

We fixed all three tolerances before any pipeline was run and before any comparison or exploratory analysis, setting each from a physical or algorithmic scale rather than by tuning it against outcomes. The period tolerance sits above the coarsest BLS grid step (about 0.1%–0.45% of the period across the 2–8 day range) yet well below the spacing between alias periods, so it neither rejects a correct period nor accepts a wrong one; the timing tolerance is half the injected duration, the scale on which a recovered mid-transit time is either usefully aligned with the true

event or not; and the depth window (±30%) matches the tens-of-percent uncertainty expected when the per-point scatter ($\approx 10^{-3}$) is comparable to the injected depth ($0.8–3\times10^{-3}$) in the high-noise trials. We made the depth band symmetric on purpose: the pipelines bias depth in opposite directions (detrend-then-search pipelines tend to under-recover, the joint model to over-recover), so a one-sided band would favour one family, whereas a symmetric band is neutral.

Rather than defend the exact values further, we validate them directly, re-grading all trials under a grid of alternative tolerances and confirming the ranking is stable (Section 3.5). This validation is a grading-only re-analysis: it reuses each pipeline's recorded recovered values and only changes the thresholds, so under the chosen tolerances it reproduces the reported pass flags exactly. A trial counts as a recovery (pass) only if all three checks succeed, and decomposing pass into its three components lets failures be attributed to period, timing, or depth rather than treated as a single opaque miss.

2.7. Statistical Analysis

Because all pipelines were evaluated on the same trials, we used paired tests. For each metric and each pair of pipelines we counted the discordant trials, those one pipeline recovered and the other did not, and applied McNemar's exact two-sided test, which under the null of equal performance treats the split of discordant trials as a fair coin and computes the exact binomial probability of a split at least as imbalanced as observed. Per-condition recovery rates are reported with 95% Wilson score confidence intervals; the Wilson interval is preferred over the standard (Wald) interval because it remains well-behaved for rates near 0 or 1, which occur in the hardest condition cells. Paired rate differences (Δ, challenger minus baseline) carry a 95% interval on the mean per-trial difference.

The primary, pre-specified comparison is P4 versus P0 on the overall pass metric; all other pipeline, metric, and per-cell comparisons are secondary. Because the secondary comparisons form a family of 16 tests (four challengers × four metrics), we control the family-wise error rate across that family with the Holm-Bonferroni step-down procedure and report the adjusted values as p_Holm in Table 1. Twelve of the sixteen comparisons remain significant at $\alpha = 0.05$ after correction: every P1, P3, and P4 comparison on every metric. The exceptions are all P2 comparisons, including its nominally significant raw result on pass ($p_{Holm} = 0.116$), so P2 is not distinguishable from the baseline on any metric once multiplicity is accounted for. Because the 2400 trials share only 30 base light curves and are therefore not fully independent, we additionally computed, for each comparison, the per-star paired rate difference and report how many of the 30 stars the challenger beats the baseline on, with an exact two-sided sign test; this confirms whether a pooled effect reflects a broad advantage or a few atypical stars.

2.8. Null-Trial False-Positive Analysis

Recovery rate measures completeness but not the false-positive rate, the tendency to “recover” a transit where none exists. We therefore generated a matched set of 2400 null trials,

identical in construction (same stars, noise, and gaps) except with no injected transit. Because a pipeline always reports some best period and depth even in pure noise, we scored every trial, real and null, by one common statistic: the recovered depth divided by the out-of-transit flux scatter at the recovered ephemeris. A common statistic is needed because the pipelines' native scores are on incomparable scales (BLS power for P0–P3, −BIC for P4); it also means the comparison depends on this choice, so it should be read as a tradeoff under one reasonable statistic rather than a definitive ranking.

2.9. Reproducibility

All trial generation and analysis used fixed random seeds via a per-star, per-condition deterministic scheme recorded in the trial index, so any trial can be regenerated exactly. All code, configuration, and the trial index are available in a public repository (https://github.com/branefungus/transit-recovery-benchmark-public). The analysis used Python 3.9 with numpy, pandas, scipy, astropy (whose BoxLeastSquares implements the search), lightkurve, astroquery, celerite2, and PyWavelets; the exact pinned package versions are listed in the repository environment file. A small number of P4 trials produced an undefined detection statistic from failed joint fits (7 of 2400 real trials and 4 of 2400 null trials) and were excluded from the false-positive analysis only; their effect is bounded in Section 3.4.

3. Results

We ran all five pipelines on the full set of 2400 trials (30 stars × 2 noise levels × 2 gap levels × 20 trials per star per cell, i.e. 600 trials in each of the four condition cells; Fig. 5 shows what these four conditions look like). The baseline P0 recovered (pass) 62.5% of trials overall. Table 1 summarizes each challenger against P0.

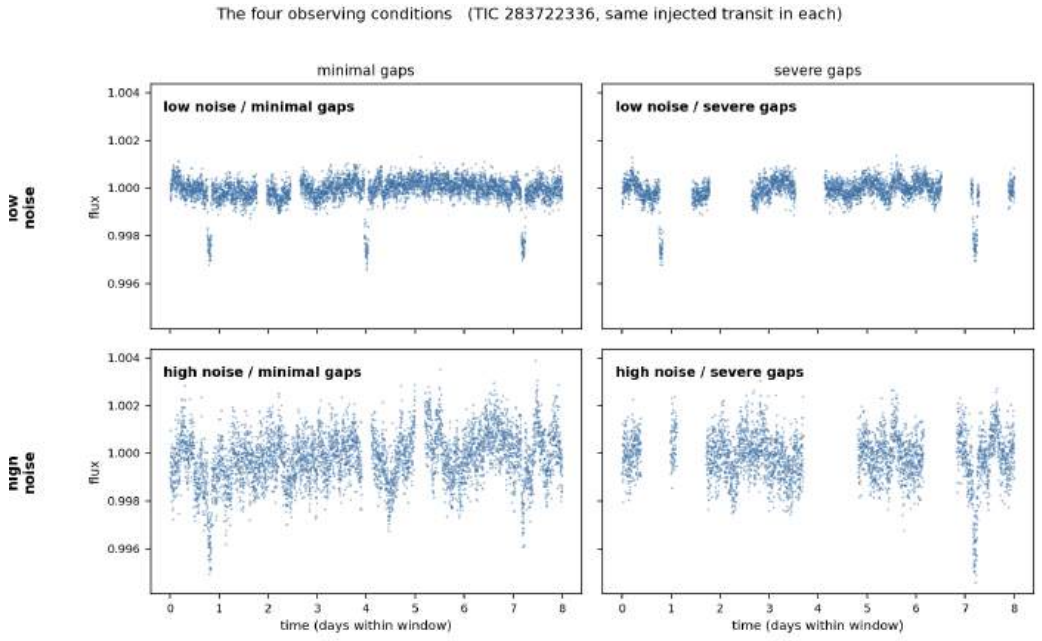


Fig 5: The four observing conditions we test, shown on the same star with the same injected transit so only the noise and gaps change. Noise increases from the top row to the

bottom; gaps increase from the left column to the right. The delta maps in Fig. 6 use this same two-by-two layout.

Table 1. Overall paired comparison of each pipeline against the P0 baseline on 2400 matched trials. Rates are the fraction of trials passing each criterion; Δ is the paired (challenger − P0) rate difference; p is the McNemar exact two-sided value; p_Holm is that value after Holm-Bonferroni correction across the family of 16 secondary comparisons. The baseline rates are pass = 0.625, p_ok = 0.640, t_ok = 0.651, d_ok = 0.736.

Pipeline	Metric	Rate	Δ (vs P0)	McNemar p	p_Holm
P1	pass	0.597	−0.029	1.5×10^{-12}	2.2×10^{-11}
	p_ok	0.611	−0.030	1.4×10^{-11}	1.8×10^{-10}
	t_ok	0.627	−0.024	8.6×10^{-8}	8.6×10^{-7}
	d_ok	0.703	−0.033	1.0×10^{-8}	1.1×10^{-7}
P2	pass	0.620	−0.005	0.029	0.116
	p_ok	0.638	−0.003	0.345	0.690
	t_ok	0.647	−0.004	0.163	0.488
	d_ok	0.733	−0.003	0.550	0.690
P3	pass	0.641	+0.016	2.8×10^{-4}	2.3×10^{-3}
	p_ok	0.660	+0.020	1.4×10^{-5}	1.3×10^{-4}
	t_ok	0.688	+0.037	3.2×10^{-17}	4.7×10^{-16}
	d_ok	0.753	+0.018	6.7×10^{-3}	0.034
P4	pass	0.643	+0.018	2.0×10^{-3}	0.014
	p_ok	0.674	+0.034	1.2×10^{-9}	1.4×10^{-8}
	t_ok	0.702	+0.050	1.3×10^{-18}	2.1×10^{-17}

	d_ ok	0. 764	+0.028	2.4×10^{-3}	0.015

3.1. Overall Recovery Relative to the Baseline

Section 3.2 and this one report pooled results, over all trials and then split by observing condition; Section 3.3 then turns to the per-star breakdown. Beginning with the pooled analysis, the joint transit model P4 and the Gaussian-process detrend P3 both improved overall recovery over P0 (pass $\Delta = +0.018$, $p = 2.0\times10^{-3}$ for P4; $\Delta = +0.016$, $p = 2.8\times10^{-4}$ for P3). The wavelet pipeline P2 was essentially neutral: its effect was indistinguishable from baseline on period, timing, and depth ($p \geq 0.16$ on each), and although its small negative effect on the combined pass metric was nominally significant before correction ($\Delta = -0.005$, $p = 0.029$), it does not survive the Holm-Bonferroni correction described in Section 2.7 ($p_{Holm} = 0.116$). The feature-regression pipeline P1 was significantly worse than baseline on every metric (pass $\Delta = -0.029$, $p = 1.5\times10^{-12}$), consistent with the possibility that the fitted trend absorbed part of the transit signal along with stellar variability.

For both P3 and P4 the improvement was largest in the timing metric (t_ok $\Delta = +0.037$ and +0.050 respectively, both $p < 10^{-16}$) and smallest in period, indicating the gains came chiefly from more accurate mid-transit time recovery rather than from finding the period, which the baseline already located well.

3.2. The P4 Advantage Is Concentrated in High-Noise Regimes

Splitting by condition shows P4's improvement is regime-dependent. On pass, P4 beat P0 by +0.030 ($p = 0.020$) and +0.033 ($p = 0.010$) in the two high-noise cells, but was not significant in either low-noise cell (−0.010 and +0.017, both intervals including zero). The timing metric shows the same pattern more sharply: $\Delta = +0.073$ and +0.085 in high noise (both $p \lesssim 10^{-8}$) versus much smaller differences in low noise. This matches the expectation that jointly modeling correlated noise helps most precisely when correlated noise is the dominant obstacle; when noise is low, simple detrending already suffices and there is little to gain.

This is a reasonable reading, but it is an interpretation rather than something the experiment proves: at least two other features of P4 could produce the same result. P4 picks among several BLS candidates using a model score, which could by itself sharpen timing in high noise where the tallest BLS peak is least trustworthy; and P4 alone searches with the limb-darkened shape matched to the injected transits (Section 2.3), which helps most where the signal is hardest to see, again the high-noise case. Separating these from correlated-noise modeling would require testing P4's pieces individually, which we did not do, so the regime dependence is a solid observation whose cause remains open. P3, by contrast, improved timing in both noise regimes, so its gains were less concentrated than P4's. The per-condition paired differences for pass are shown as delta maps in Fig. 6.

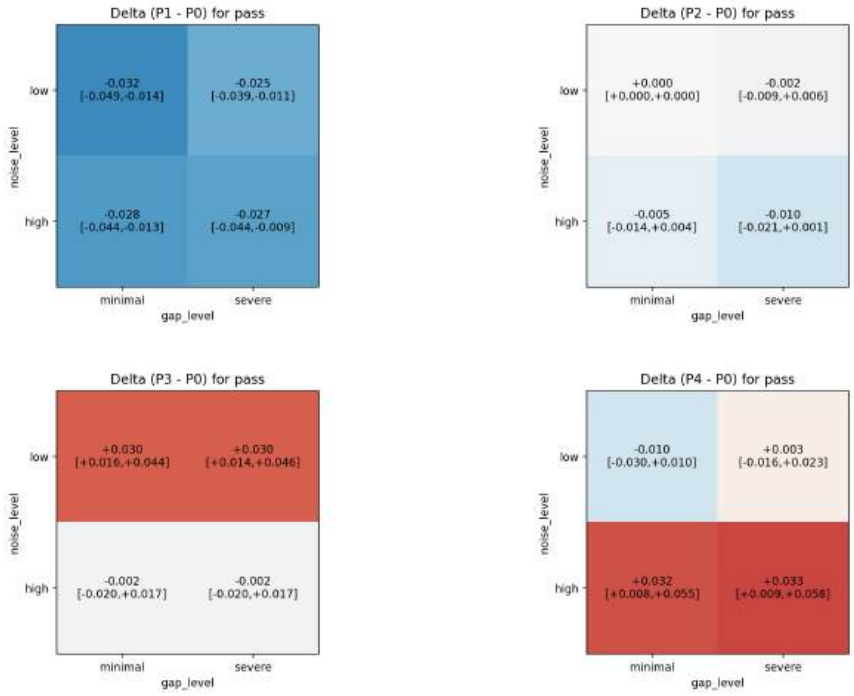


Fig 6: Paired difference in overall recovery rate (challenger minus P0) for the pass metric, by noise level (rows within each panel) and gap level (columns within each panel), with 95% intervals. Red indicates the challenger outperforms P0, blue the reverse. All four panels show the overall pass metric only. On that metric P1 (top left) is uniformly worse; P2 (top right) is near zero everywhere; P3 (bottom left) improves in the low-noise row; and P4 (bottom right) improves in the high-noise row. The pattern differs by metric: as reported in Section 3.2, P3 improves mid-transit timing in both noise regimes, so its concentration in the low-noise row here is specific to the combined pass criterion and should not be read as a statement about timing.

3.3. P4's Advantage Holds Across Individual Stars; P3's Does Not

Because the 2400 trials share only 30 base light curves, we checked whether the pooled advantages reflect broad effects or a few atypical stars. P4 beat P0 on 19 of 30 stars, lost on 7, and tied on 4 (exact sign test $p = 0.029$), confirming that its advantage is broad across the sample. By contrast, although P3 improved in the pooled, trial-weighted analysis, it won on only 7 of 30 stars and lost on 14 (sign test $p = 0.19$, not significant): its overall gain is driven by large improvements on a minority of stars rather than a consistent per-star advantage. P1's disadvantage was broad (lost on 16 of 30, $p = 0.004$) and P2 was mostly tied (20 of 30 stars), consistent with their pooled results. This distinction (P4's advantage is more consistent across individual stars than P3's) is visible only in the per-star view and is a central qualification of our findings.

3.4. Detection Against Noise: A Completeness-False-Positive Tradeoff

The recovery analysis above measures completeness: how reliably a pipeline recovers a transit that is present. It does not measure how well a pipeline tells a real transit from pure noise.

To assess that, we scored the real trials and the matched null trials (Section 2.8) by the common depth-over-scatter statistic and swept a threshold, tracing completeness (real trials that pass recovery and clear the threshold) against false-positive rate (null trials that clear it). The curves are shown in Fig. 7. As noted there, this comparison depends on the chosen statistic and should be read as a tradeoff under one reasonable choice rather than a definitive ranking.

Under this statistic, the four box-search pipelines (P0–P3) behave similarly, reaching roughly 50–55% completeness at a 10% false-positive rate and about 23% at a 5% rate. P4, despite achieving the highest unconditioned recovery rate (Table 1), separates real transits from noise less effectively: its completeness at a fixed false-positive rate is consistently lower (about 34% at 10% and 11% at 5%), placing its curve below the others in the low-false-positive regime that matters most for candidate vetting. In other words, P4's advantage in recovering a transit that is present does not translate into a better statistic for deciding whether a transit is present at all; the simpler box-search pipelines discriminate against noise at least as well.

A small number of P4 trials gave an undefined statistic when the joint fit failed and were excluded here: 7 of 2400 real trials and 4 of 2400 null trials. These cannot explain P4's lower curve. The dropped null trials are at most 4/2400 = 0.17% of the null set, so even if all had counted as false positives at every threshold the false-positive rate could move by only 0.17 percentage points; and all 7 dropped real trials had already failed recovery, so keeping them would only lower P4's completeness further. These are genuine, if rare, failures of the joint fit rather than a neutral loss of data.

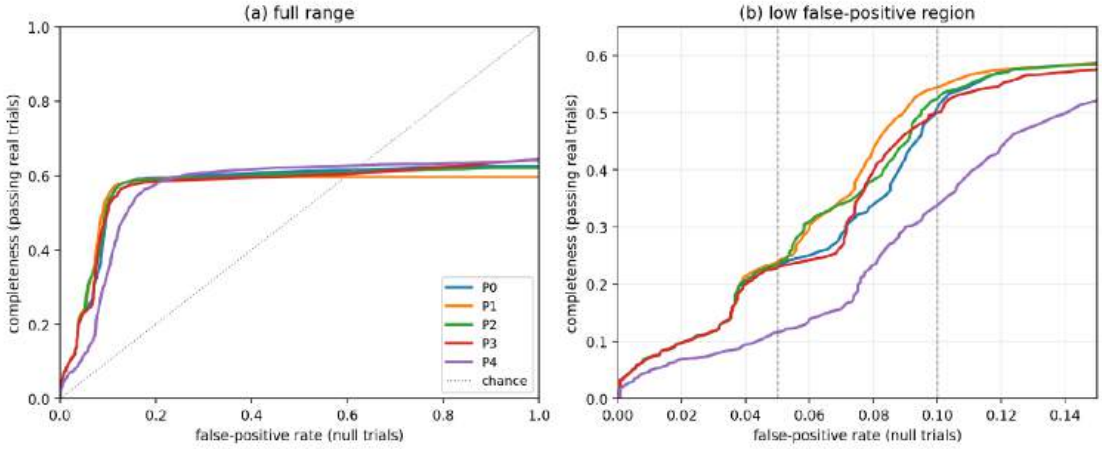


Fig 7: Completeness (fraction of real transits recovered) versus false-positive rate (fraction of injection-free null trials exceeding the detection threshold), under a common depth-over-scatter detection statistic swept across thresholds. Panel (a) shows the full range; panel (b) magnifies the low-false-positive region below 15%, where candidate vetting operates and where the separation between the pipelines is clearest. P0–P3 behave similarly; P4 lies below them in the low-false-positive region despite its higher unconditioned recovery rate. Dashed lines in panel (b) mark the 5% and 10% false-positive rates quoted in the text.

3.5. Sensitivity to the Recovery Tolerances

The pass result depends on three cutoffs (how close in period, timing, and depth a recovery must be to count). To check that our conclusions are not an accident of the specific cutoffs we chose, we re-graded all 2400 trials under a grid of alternative cutoffs: the period tolerance at 0.5%, 1%, and 2%; the timing tolerance at 0.5, 1, and 1.5 hours; and the depth window at several widths from tight (0.85–1.20×) to loose (0.60–1.60×), including a narrower symmetric band. Each setting changes only the named cutoff, holding the other two at their chosen values. The re-grading reuses each pipeline's already-recorded recovered values, so only the cutoffs change, not the pipelines' outputs, and under the chosen cutoffs it reproduces the reported pass rates exactly. The relative order of the four detrend-then-search pipelines, P3 > P0 > P2 > P1, and P3's significant lead over the baseline held up under every setting we tried, including the tightest and loosest cutoffs. The absolute pass rates moved only slightly under the period and timing changes (by about a percentage point or less), and the ranking among these four never changed.

P4 is the one pipeline whose standing depends on the cutoffs, and it is sensitive to two of them, in a way that is revealing rather than worrying. Under the depth window we use and any wider one, P4 leads; but under a much tighter depth window (for example 0.85–1.20× the true depth) P4 falls behind the baseline (its lead over P0 reverses to $\Delta = -0.028$, $p = 1\times10^{-4}$). It also slips under a much tighter timing window: at a half-hour tolerance P4's lead over P0 shrinks to nothing ($\Delta = -0.004$, $p = 0.54$) while P3 stays ahead. Both sensitivities have the same root cause. P4 gets depth wrong in the opposite direction from the other pipelines, its joint fit tends to over-estimate depth, so more of its trials land above the upper depth limit than below the lower one, whereas P0–P3 tend to slightly under-estimate it; part of P4's advantage therefore comes from depths landing in the upper part of the allowed window, which a tighter window removes. Under a tighter timing cut, P4's candidate selection loses the edge it holds when timing is judged loosely. In short, P4's lead in parameter recovery is real under our tolerances and looser ones but is contingent on the depth and timing cutoffs not being made much stricter, whereas the P3 lead carries no such dependence.

4. Discussion

4.1. Summary of Findings

Of four alternatives to a simple rolling-median baseline, two improved transit recovery, one was neutral, and one was harmful. The joint trend-plus-transit model (P4) and the Gaussian-process detrend (P3) both raised the overall recovery rate over the baseline P0; the wavelet pipeline (P2) was essentially neutral; and the feature-regression pipeline (P1) was significantly worse on every metric. For both P3 and P4 the gains were largest in mid-transit timing. Two qualifications matter: P4's advantage was concentrated in the high-noise cells and was broad across individual stars (19 of 30), whereas P3's pooled gain did not hold star by star (7 of 30). Throughout the discussion that follows, “P4's advantage” should be read as an advantage of its whole architecture, joint fitting, model-based candidate selection, and a matched transit template together, since these remain confounded and we did not isolate them (Section 3.2). The remainder of this section considers why these patterns arose.

4.2. Mechanisms Behind the Observed Patterns

Why the gains showed up in timing, not period. The baseline already found the period well, leaving little room to improve. The period comes from the spacing between transits, which BLS measures from all events together, so it is forgiving of any single dip's shape. Mid-transit time is more delicate: it depends on where the bottom of each dip falls, and a trend sloping across a transit shifts that apparent bottom without changing the spacing. Trend-removal errors therefore show up mostly in timing, which is where a better trend model helps most. The same reasoning explains the smaller depth gain: a trend through the transit distorts how deep it looks while barely touching the period.

Why P3 helped some stars but not others. Almost all of P3's gain came from a few light curves where the baseline did very badly: one star's recovery rate jumped from 0.09 to 0.58, another from 0.00 to 0.20, and the three biggest improvements account for about 91% of P3's per-star gains. On the rest it was flat or slightly worse. The Gaussian process seems to help most when a star's variability is strong and structured enough to fool a rolling median, and can slightly hurt otherwise by following noise a median would smooth over. This is why the pooled result, dominated by the few stars with the most room to improve, and the star-by-star result disagree for P3.

Why gaps mattered less than noise. Averaged over gaps, going from low to high noise dropped the baseline recovery rate by 0.31 (0.78 to 0.47); averaged over noise, going from minimal to severe gaps dropped it by only 0.04 (0.64 to 0.61), about eight times less. Even severe gaps remove only about 12% of the points, and BLS folds whatever remains over a multi-week baseline, so several transits usually survive. Correlated noise is worse because it does not merely remove data but mimics a transit, since we tuned it to the transit timescale: missing data mainly costs sensitivity, whereas transit-like noise costs sensitivity and reliability, and it is the latter that the recovery criteria penalize.

What stellar properties predict performance. Our study cannot answer this directly, since we picked stars for brightness and uniformity rather than to span stellar types. The results are suggestive, the stars where pipelines differ most have the lowest baseline recovery rates, where trend removal is hardest, but relating per-star performance to variability, rotation, and crowding is a task for a differently designed sample.

4.3. Recovery Versus Detection

A pipeline that recovers transit parameters well is not necessarily the one that best distinguishes transits from noise. Our null-trial analysis made this explicit: P4's recovery advantage did not extend to noise discrimination, where it fared no better than the simpler box-search pipelines. Completeness and false-positive resistance are therefore distinct axes of performance, and the best pipeline depends on the goal: for characterizing a transit believed to be present, especially under heavy correlated noise, the joint model is advantageous; for vetting candidates against false alarms, the simpler approaches are at least as good by our measure.

4.4. Practical Recommendations

We separate two kinds of takeaway: recommendations specific to the pipelines as tested here, and a more general methodological lesson.

Benchmark-specific (Table 2 summarizes these). P4 (the joint-architecture pipeline) is favored for parameter recovery under heavy correlated noise, but offers no advantage in low noise and is no better at rejecting false alarms, where the simple baseline (P0) is at least as good and far cheaper. The Gaussian-process detrend (P3) is a reasonable alternative for recovery, with the caveat that its gains do not hold across most individual stars. The feature-regression pipeline (P1) is not recommended. These statements are established only for the bright, confirmed-planet sample and the fixed transit family used here.

General lesson. Independent of which pipeline wins, the study shows that recovery accuracy and false-positive resistance are distinct axes that a single benchmark can rank differently, so a pipeline evaluation that reports only one of them can mislead. We would encourage future comparisons to report both, as we do here.

Table 2. Recommended pipeline by goal and noise regime, based on this study. "Characterize" means recovering the parameters of a transit believed to be present; "vet" means distinguishing a candidate from noise. Recommendations follow the recovery results (Table 1, Fig. 6) and the false-positive analysis (Fig. 7).

Goal	Noise regime	Recommended pipeline
Characterize a transit	high (correlated) noise	P4 (joint model)
Characterize a transit	low noise	P0 (baseline); P4 gives no gain
Vet candidates vs. false alarms	either	P0 (baseline); simple and reliable

4.5. Limitations

Several limitations bound these conclusions.

The input data has already been partly cleaned. We inject into PDCSAP flux, from which the mission has already removed instrumental effects, so the pipelines are judged on how well they remove what remains, mainly stellar variability and our added correlated noise, not raw instrumental signals. The conclusions therefore apply to the common case of cleaning already-processed light curves, not raw data, and a method whose strength is removing instrumental effects is at a disadvantage here. This also shows up in the ranking: P1's segment-offset terms target step-like jumps that the mission processing has largely already removed, which probably contributes to its poor showing.

The stars are not a representative sample. The 30 targets are the brightest confirmed-planet TOI hosts, chosen for data quality. Fainter, more variable, or more crowded stars are harder to detrend, so both the recovery rates and the ranking we find here may not carry over to a typical, fainter sample.

P4 differs in more than its detrending. P0–P3 are two-step (detrend-then-search) pipelines that differ in a single step, but P4 also selects among candidate solutions using a model score and searches with a matched transit shape. Its advantage should be credited to the joint approach as a whole, not to correlated-noise modeling specifically, since we did not test its pieces separately (Section 3.2).

The conclusions depend on the recovery cutoffs. The sensitivity check (Section 3.5) shows the ranking of the two-step pipelines (P0–P3) is stable across every cutoff we tried, but P4's lead over the baseline, while solid under our cutoffs and looser ones, shrinks or reverses under a much tighter depth or timing window, because P4 tends to over-estimate depth and because part of its edge sits in the loosely-judged tail of the timing distribution.

The injected and recovered signals come from the same family. Every pipeline searches for, and every trial contains, a transit from one closely specified model: the same limb-darkening law, impact parameter, and duration, with the depth template matched to the injected shape by design (Section 2.3). No pipeline is ever confronted with a signal whose shape it does not already expect, which makes the benchmark structurally favorable. A real search faces transits that depart from any single template, so the ranking here should be read as holding for the matched-family case; how the pipelines reorder when the injected and searched families diverge (for example, grazing or strongly mismatched limb darkening) is what the robustness test of Section 4.7 would probe.

A few narrower limitations remain. Fixed limb-darkening coefficients were used for every star. The pipelines were given the true duration and impact parameter, so recovery rates should be read as upper bounds relative to a fully blind search. The trials span only 2–8 day periods, a single 2-hour duration, and a modest depth range, so the results should not be extrapolated to long-period or grazing transits. And the false-positive comparison depends on the chosen common statistic, as discussed above.

4.6. Future Work

Natural extensions include widening the period, duration, and depth ranges and adding grazing geometries; testing robustness of the P4 result to deliberately mismatched limb-darkening coefficients, to separate a genuine modeling advantage from a matched-template effect; comparing alternative common detection statistics to test the robustness of the false-positive ranking; enlarging the stellar sample; and packaging the injection-recovery framework so that other pipelines can be benchmarked against the same trials.

5. Conclusion

We benchmarked five transit-detection pipelines that share a common periodic search and differ in how they remove stellar trends, using a controlled injection-recovery experiment on 30 real TESS light curves across a 2×2 grid of noise and gap severity. Relative to a rolling-median baseline, a joint trend-plus-transit model and a Gaussian-process detrend improved recovery, a wavelet pipeline was neutral, and a feature-regression detrend was worse. P4's advantage, attributable to its joint architecture as a whole rather than to joint modeling alone, was broad

across stars and concentrated in high noise, where it most improved mid-transit timing. A null-trial analysis, however, showed that this recovery advantage did not translate into better separation of transits from noise. No single pipeline is therefore best for every purpose: the joint model was preferable for recovering parameters under high correlated noise, while simpler box-search pipelines were at least as effective for vetting candidates against false alarms. More broadly, recovery accuracy and false-positive resistance are distinct axes of performance that should be reported together.

Data and Code Availability

All photometry is public, obtained from the Mikulski Archive for Space Telescopes (MAST) through Lightkurve. A machine-readable table in the repository lists the 30 stars and their exact SPOC products with direct MAST download URLs, so the identical input light curves can be retrieved. All analysis code, the configuration file (defining every tolerance, injection range, noise and gap setting, and random seed), the trial index, and the per-pipeline results tables are in a public repository (https://github.com/branefungus/exoplanet-transit-detection-pipelines/tree/main).

Acknowledgments

All implementation, experimentation, and analysis were carried out by the student researcher; no mentor-run processing was used for the reported results. I thank Dr. Ana-Maria Piso-Grigore for advice on aspects of the methodology and for her feedback on the manuscript. This work made use of data from the TESS mission, obtained from the MAST archive, and of the open-source packages cited above.

Works Cited

Astropy Collaboration, et al. "The Astropy Project: Building an Open-Science Project and Status of the v2.0 Core Package." *The Astronomical Journal*, vol. 156, 2018, art. 123.

Barclay, Thomas, Joshua Pepper, and Elisa V. Quintana. "A Revised Exoplanet Yield from the Transiting Exoplanet Survey Satellite (TESS)." *The Astrophysical Journal Supplement Series*, vol. 239, 2018, art. 2.

Christiansen, Jessie L., et al. "Measuring Transit Signal Recovery in the Kepler Pipeline. I. Individual Events." *The Astrophysical Journal Supplement Series*, vol. 207, 2013, art. 35.

Christiansen, Jessie L., et al. "Measuring Transit Signal Recovery in the Kepler Pipeline. II. Detection Efficiency as Calculated in One Year of Data." *The Astrophysical Journal*, vol. 810, 2015, art. 95.

Christiansen, Jessie L., et al. "Measuring Transit Signal Recovery in the Kepler Pipeline. III. Completeness of the Q1–Q17 DR24 Planet Candidate Catalog." *The Astrophysical Journal*, vol. 828, 2016, art. 99.

Foreman-Mackey, Daniel, et al. "A Systematic Search for Transiting Planets in the K2 Data." *The Astrophysical Journal*, vol. 806, 2015, art. 215.

Foreman-Mackey, Daniel, et al. "Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series." *The Astronomical Journal*, vol. 154, 2017, art. 220.

Hippke, Michael, and René Heller. "Optimized Transit Detection Algorithm to Search for Periodic Transits of Small Planets." *Astronomy & Astrophysics*, vol. 623, 2019, art. A39.

Hippke, Michael, Trevor J. David, Gijs D. Mulders, and René Heller. "Wōtan: Comprehensive Time-Series Detrending in Python." *The Astronomical Journal*, vol. 158, 2019, art. 143.

Kipping, David M. "Efficient, Uninformative Sampling of Limb-Darkening Coefficients for Two-Parameter Laws." *Monthly Notices of the Royal Astronomical Society*, vol. 435, 2013, pp. 2152–2160.

Kovács, Géza, Shay Zucker, and Tsevi Mazeh. "A Box-Fitting Algorithm in the Search for Periodic Transits." *Astronomy & Astrophysics*, vol. 391, 2002, pp. 369–377.

Kovács, Géza, et al. "Optimal Trend Filtering and Transit Search of K2 Data." *Astronomy & Astrophysics*, vol. 596, 2016, art. A48.

Lightkurve Collaboration, et al. "Lightkurve: Kepler and TESS Time Series Analysis in Python." *Astrophysics Source Code Library*, 2018, ascl:1812.013.

Luger, Rodrigo, et al. "EVEREST: Pixel Level Decorrelation of K2 Light Curves." *The Astronomical Journal*, vol. 152, 2016, art. 100.

Mikulski Archive for Space Telescopes (MAST). *TESS Mission Data Products*, accessed via Lightkurve. Space Telescope Science Institute, https://mast.stsci.edu/.

Morris, Brett M., et al. "The Stellar Variability Noise Floor for Transiting Exoplanet Photometry." *The Astronomical Journal*, vol. 160, 2020, art. 5.

Pont, Frédéric, Shay Zucker, and Didier Queloz. "The Effect of Red Noise on Planetary Transit Detection." *Monthly Notices of the Royal Astronomical Society*, vol. 373, 2006, pp. 231–242.

Taaki, Jamila S., Farzad Kamalabadi, and Athol J. Kemball. “A Search for Exoplanet Candidates in TESS 2-Minute Light Curves Using Joint Bayesian Detection.” *The Astronomical Journal*, vol. 170, 2025, art. 20.

Soft-Core Radicalization: Feminism and the Digital 'Trad-Wife' Aesthetic in Emergent Conservative Gen-Z Politics By Sainzaya Buyankhuu

Abstract

This paper analyzes the 'trad-wife' aesthetic as a digital performance of domesticity and its influence on emergent Gen-Z political ideologies. Content from popular trad-wife influencers and reactionary Gen-Z responses are examined to trace how lifestyle branding embeds conservative values. Using examples from TikTok and Instagram, two of the most popular trad-wife creators are analyzed through their most-viewed videos, while reactionary content such as duets, parodies, and critical commentary provides evidence of generational reception. This method defines soft-core radicalization as the subtle circulation of conservative ideals through seemingly apolitical or aestheticized media. The research found that the trad-wife aesthetic commodifies domesticity, embeds conservatism within femininity, and provokes both critique and defense framed through choice feminism. Theorizes that this aesthetic functions as soft-core radicalization, shaping Gen-Z political consciousness by masking ideological hierarchies under personal lifestyle branding.

Keywords

Trad-Wife, Feminism, Digital Aesthetics, Conservatism, Politics, Gen-Z, Domesticity, Soft-core Radicalization, Femininity, Influencers, Marketability, Gender Roles, Social Media, TikTok, Instagram, Personal Branding, Political Ideology, Commodification, Nationalism, Soft Power

Introduction

In late 2023 to early 2024, Nara Smith went viral on Tiktok and other social media platforms by cooking food for her children from scratch and embodying the 'trad-wife' lifestyle. Accumulating up to 3 million followers (and counting) since she began posting content, according to an article published by People magazine (Aizin). The 'trad-wife' aesthetic, aka the traditional wife is defined as a married woman who embraces traditional gender roles and values. Women who embody these values online have been consequently deemed as trad-wife influencers and build personal brands centered around lifestyle content. The primary consumers of this content, or the most 'influenced' demographic, are Gen-Z, particularly young women. However, the impact of this cultural phenomena can be observed from the online audiences of these trad-wife influencers on social media platforms, particularly TikTok, Instagram, and Pinterest as platforms of ideological lifestyle branding.The trad-wife aesthetic, being derived from the values of the traditional wife, serves as a curated online performance of traditional domestic femininity, often tied to anti-feminism, nationalism, or conservative values. At first glance, the "trad wife" may appear to be an innocuous lifestyle choice, filled with curated images

of home cooking, modest clothing, and family centered ideals. However, the aesthetic holds significant political weight through embedding values of gender, nation, and tradition into a digital performance.The power is not solely derived from imagery itself, but the feelings it invokes; coziness, stability, belonging and the communities it fosters in online spaces.

The traditional wife emerged following the Cold War in the United States. By 1948, after years of depression and violence, economic and societal stability were desperately desired by the middle class. The importance of the appeasement of the middle class (and the subsequent rise of the stay-at-home, traditional, wife role) was highlighted in the 'Kitchen Debate'. A spirited conversation held between former US Vice President Richard Nixon and Premier Nikita Khrushchev over the merits and flaws of capitalism and communism. Their views aligned on the meaning of success: big houses, cool appliances and pointedly-consumer citizenship (Rogers-Cooper). This conversation has since been dubbed as a synopsis of post-Cold War politics in the United States as a promotion of capitalism. Spurring domesticity campaigns from the federal government and popular media to encourage women to resort back to homemaking, "fueled by a newly electrified, scientific kitchen and appliances, defining a women's proper role as that of 'home manager'" (Takeuchi). After many women took on male-dominated professions amidst the war out of necessity, post Cold War meant a new era for the United States (Babb). As a result of these campaigns, marriage and birth rates skyrocketed and women were enthralled in the duties of housewives and mothers. The resurgence of the trad wife has caused a stir amongst online circles for its religious connotations, with it being popularized in evangelist and fundamentalist circles for its accordance with the values of Christianity. While not all women subscribe to the definition of what it means to be a trad-wife, the underlying Christian values of the community are prevalent (Keegan). The trad-wife aesthetic now reappears as curated digital content blending homemaking, vintage femininity, and ideological radicalization commodifying conservative values (Sykes and Hopner).

This research paper aims to prove the greater implications of the trad-wife aesthetic on the younger generation and US politics. Social media's accessibility and prevalence in modern society lends itself to influence a wider sphere of viewers than traditional forms of media. Generation-Z, or Gen-Z, are people who were born between 1997 to 2012 and are characterized by their upbringing being amidst the digital age, growing up with social media and the Internet (Atay and Ashlock). Gen-Z's proximity to social media makes them the ideal demographic to analyze the longer term effects of the digital aesthetic, the trad-wife.

The historical evidence of the traditional wife is rooted in anti-feminism. Amidst the 1950s women had already gained the right to vote; however, were still living under the guise of societal expectations. The second wave of the women's liberation movement started not long thereafter. Radical feminists challenged the traditional notion of politics

and prioritized women's experiences at the forefront of the movement led with the slogan "the personal is political" (Nachescu). Highlighting how women's lived experiences translate into a growing societal consciousness and political reforms towards gender equality. The fight for women's equality has led to immense progress, warranting women autonomy in personal and political endeavors. Repopularizing the trad-wife amongst younger generations on influential social media platforms aligns with oppressive anti-feminist rhetoric that coined the term in the post Cold War Era. The archetype now exists in liminality, between the dated conservative values and romanticized labor (Proctor). Leading to the research question guiding the following paper, How does the rise of the 'trad-wife' as a curated aesthetic shape and reflect Gen Z's political ideologies? This research will offer insight into the effects of digital content on the new demographic of American voters, what it implies about the future of politics and how to approach soft-core radicalization content with nuance.

Methodology

Content of popular trad wives from various social media platforms will be analyzed as evidence. To do this, I will use a search engine to determine the two most popular trad-wife influencers, or those with the highest number of followers across various social media platforms (i.e Tiktok, Instagram, Pinterest). It should be noted that influencers that are recognized for making trad-wife content don't necessarily identify with the label. For the criteria of determining whether or not a creator falls under the trad-wife archetype, they must be a woman known for making traditional home-making content (i.e cooking from scratch, cleaning, taking care of children, etc.) and be recognized by their viewership as falling into the trad-wife aesthetic. This includes appearing on social media platforms under '#tradwife', news articles, and comments under their videos labelling them in this archetype. From these two content creators I will then analyze a piece of content they've posted. The video from each content creator will be chosen by meeting the following conditions: the video may not be a paid sponsorship, the video must be one of the most viewed/interacted with videos from their page, it must accurately encapsulate their personal brand and not be satirical.

Reactionary content created by Gen-Z, such as parodies, duets, and critical commentary, provides important insight into how younger audiences interpret the trad-wife aesthetic. Many of these responses exaggerate or mock domestic roles, such as cooking elaborate meals or serving husbands, in order to expose the conservative values embedded in trad-wife content. This practice reflects what Proctor (2022) identifies as the tension between trad-wives' use of feminist rhetoric and the anti-feminist structures their content promotes (Proctor). By ridiculing or critically responding, Gen-Z creators demonstrate an awareness of how lifestyle branding can mask ideology, echoing Nachescu's (2009) reminder that the personal is always political (Nachescu). At the same time, defenders of trad-wife content often fall back on "choice

feminism," which Ferguson (2010) critiques as a framework that excuses systemic inequality by framing individual decisions as empowerment. These reactionary posts show that digital resistance functions as both satire and political discourse, allowing Gen-Z to question the influence of conservative gender norms within online culture.

Performing Tradition: The 'Tradwife' as a Digital Aesthetic

The following TikTok video, posted on May 18, 2025, features two extremely popular 'trad-wife' influencers: Nara Smith (seen in a white polk-dotted dress) and Hannah Neelemen (facing away from camera in floral patterned dress). Nara Aziza Smith, known as Nara Smith online, is a 23 year old content creator on Tiktok and Instagram. She's an ex model and is married to a 27 year old model Lucky Blue Smith with three children. Nara began posting videos in 2023 and has since amassed a following of 17 million across social media platforms. She's known for cooking meals for her family in lavish outfits and whispering in her Tiktoks. Hannah Neelemen, known as Ballerina Farm online, is a 35 year old content creator on Tiktok and Instagram and has 20.8 million followers across Tiktok and Instagram.The name, Ballerina Farm, comes from her passion of dance as a Julliard trained dancer. She's married to a 35 year old Daniel Neelemen and they have 8 children.The two are small business owners, selling products produced from their farm on their website, ballerinafarm.com. Hannah's content is centered around home cooked meals, managing farm chores, and promoting products from their business.

Figure 1. [Nara Smith and Hannah Neilman]

This particular video showcases the two preparing a multi course dinner for their children and husbands. Composed of spaghetti and meatballs, roasted rhubarb and ice cream, asparagus, salad, sourdough bread and garlic herb butter, each component made meticulously from scratch. Multiple clips are shown of the two women preparing each part of the meal, and towards the beginning and end of the video the husbands and children are shown. The two husbands are shown talking off towards a truck, each holding a baby, their children. At the end of the video, the children are seen plating their own food from the various dishes made, Nara holds her child and caresses their head, and the husbands take bites of their food. Hannah does the voiceover, describing the dishes made and how wonderful it tasted and the video is captioned with, “Spaghetti has always been our kids favorite meal. What’s your go-to family meal?”. While the video doesn’t overtly state ‘trad-wife’, it reflects it through its contents. The act of the two women preparing dinner for their families feels habitual, traditional even. Within their family dynamic, and reflected in their other content, it’s understood that the wife is primarily handling the cooking. The husbands are entirely exempt from this task; however, they are seen taking care

of their children for a small portion of the video, as to relieve the women's burden and consequently un-adhering to gender roles. The lack of labeling the content under the #tradwife, may make it hard to ascertain the accordance to conservatism and anti-feminism. However, the large viewership of content reflecting the characteristics of this familial dynamic maladapts to inexplicitly radicalize conversativism through 'femininity'(Zahay). Femininity masking the inherently politicized representation of tradition.

> "The "momosphere"—a collection of parenting, relationship, cooking, and crafting blogs and social media—has seen exponential growth during the socially distant pandemic years. It has also seen the rise of a new domestic online persona:the 'tradwife.'...To combat the inevitable critiques of misogyny, many self-identified trad wives use feminist rhetoric to frame the movement as a choice they are making about how to live their own lives as empowered women" (Proctor).

The video romanticizes household chores through natural lighting, well dressed wives, and a voice-over emphasizing the positive aspects of household chores. Influencing their audiences into a conservative lifestyle, especially young impressionable female audiences. To viewers, domesticity appears to be the path leading to life as picturesque and Instagram-able as Nara and Hannah's. The reception of the video, however, told a different story. Despite the high interaction this particular video got in relevance to the others on their respective Tiktok pages, comments reflect the disdain for its contents. One user mentioned Nara being 23 years old, noting the oddity of her submission to motherhood as a young adult. Another comment says that "she should be at the club" on similar lines. Viewers are hypercritical of this one post in particular. The disproportionate amount of viewers may have resulted in the algorithm showing the video to people outside Nara or Hannah's usual fanbase. On the other hand, some in the comment section rise to Nara and Hannah's defense, by saying things along the lines of "Why can't you just let them be happy?" and "They aren't being forced into doing this. It's their choice."

While these viewers come from a place of good intention, the notion that these trad-wives are relinquished of critique for projecting anti-feminism because they choose this way of life is choice feminism. Choice feminism is adhering to aspects of feminism and disregarding others from a fear of being too political. It allows feminists to side-step making the personal political and doesn't enact change to the status quo (Ferguson). Making the personal political allows people to recognize the systemic systems in place that lead to the issues within people's personal lives. The gender roles that dictate the distribution of household chores for example. It allows an individual's biases to be examined and the growth of the collective consciousness when personal conflict arises (Rogan and Budgeon).

The performance of tradition in Nara and Hannah's video is not simply about cooking dinner but about embodying a curated ideal of womanhood that resonates with cultural nostalgia. Judith Butler's (1990) notion of performativity is useful here, as the repeated acts of homemaking, child-care, and marital devotion become a performance that naturalizes gender roles as timeless truths (Butler). In this context, "tradition" does not merely describe inherited practices but a consciously staged representation that situates domestic labor within an aesthetic of stability, morality, and femininity. For Nara, the whisper-toned narration and polished presentation transform household work into a performance of delicacy and grace, while for Hannah, the pastoral backdrop of her farm frames homemaking as rooted in heritage and authenticity. These representations conceal the constructed nature of tradition by presenting it as organic, while in reality they are carefully mediated digital performances that commodify nostalgia for conservative ideals. In this way, both creators have made tradition a performance, utilizing its cultural weight to attract audiences while embedding ideological significance into

ordinary domestic tasks.

Commodifying Domesticity: The Marketability of the 'Trad-Wife'

A defining feature of the trad-wife aesthetic is the merging of domestic labor with marketable content. For example, Hannah Neeleman's Ballerina Farm brand demonstrates how homemaking becomes both a lifestyle and a consumer product. Through curated videos of breadmaking, farm chores, and child-care, Neeleman projects an image of idyllic family life while simultaneously advertising her farm-raised meat and homemade goods. The rustic mise-en-scène of sunlit kitchens, handmade bread, vintage aprons, creates an aura of authenticity that masks the commercial motives of her brand.

This blending of home life and market aligns with what Sykes & Hopner (2023) describe as the commodification of conservative values, where tradition is packaged and circulated as consumable media (Sykes and Hopner). What appears to be genuine homemaking is, in reality, a curated and monetized performance of femininity. The trad-wife figure is rebranded as both mother and entrepreneur, sustaining traditional gender roles while cloaking them in capitalist productivity. This reflects the legacy of the Cold War-era "kitchen debate" and subsequent domesticity campaigns (Rogers-Cooper; Takeuchi), where women were positioned as "home managers" whose domestic labor was reframed as central to the nation's success.

For Gen-Z, this contradiction is especially evident. Some followers embrace the aesthetic as authentic and aspirational, while others critique it as a commodified fantasy. As Zahay (2022) notes, femininity often functions as a mask for political ideology, embedding conservatism within seemingly innocuous content (Zahay). Far from subverting patriarchy, the monetization of homemaking reinforces traditional hierarchies while making them more palatable through capitalist consumption. The trad-wife thus emerges not only as a cultural aesthetic but also as a profitable brand strategy that circulates conservative ideals under the guise of wholesome domesticity.

Nara Smith's TikTok account shows an engagement rate of about 8.82% with around 11.8 million followers, producing videos that average roughly 4.5 million views each. Hannah Neeleman's Ballerina Farm exhibits similarly strong performance, with an Instagram engagement rate of 1.73% and average likes in the hundreds of thousands per post. Their audience is largely composed of Gen-Z young women seeking comforting lifestyle content. Both creators use strategically curated platform tools: trending audio or voiceovers, aspirational visuals, and collaborations to achieve algorithmic reach. Their content rarely references politics explicitly, instead embedding conservative-aligned ideals through aesthetics of home, caregiving, and domestic bliss. By relying on soft power rather than overt political rhetoric, they build large followings with limited backlash and high audience trust.

Gen-Z Critique and Political Reflection

Survey data suggest that Gen-Z is not uniformly more conservative than previous generations, though there are signs of shifting dynamics. According to a 2024 poll by PRRI, only 28 percent of Gen-Z adults identify as conservative, a notably lower amount than among older groups, while 43 percent identify as liberal, more than any other generation ("A Political and Cultural Glimpse Into America's Future"). However, younger Gen-Z members, ages 18 to 24, are reporting increasing openness to conservative ideas. Reporting from *Vox* highlights that this subgroup shows greater support for Republican policies and candidates, especially among men, which marks a divergence from older Gen-Z members who remain more solidly progressive (Paz). Thus, while Gen-Z as a whole remains more liberal on many social issues compared to previous generations, the popularity of trad-wife aesthetics, fueled by soft power and aesthetically mediated content, can tap into the more conservative leanings of certain subgroups. This contributes to Gen-Z's political heterogeneity, where some embrace progressive values while others are drawn to traditional or conservative ideals when presented through lifestyle content rather than overt political appeal.

Although trad-wife influencers have amassed millions of followers, their content has also prompted significant reactionary discourse, particularly from Gen-Z audiences. On TikTok, duets, parodies, and critical commentary expose the political undertones of the trad-wife aesthetic. Many young creators exaggerate domestic roles in parody, mocking captions like "serving my husband" or "submission is love," often inserting critiques about autonomy and feminism. This response reflects what Proctor (2022) identifies as the tension between trad-wives' use of feminist rhetoric ("it's my choice") and the inherently anti-feminist structures their lifestyle perpetuates (Proctor). These responses parallel historical depictions of feminist thought. As Nachescu (2009) emphasizes, second-wave feminists challenged the notion that personal choices could be separated from politics, leading to the slogan "the personal is political (Nachescu)." When Gen-Z users resist trad-wife content, they continue this tradition by exposing how domestic aesthetics perpetuate systemic gender inequality. By contrast, defenders of trad-wife influencers often adopt a "choice feminism" stance (Ferguson), arguing that because women voluntarily choose this lifestyle, it should be exempt from critique. Yet, as Hanisch (2006) notes, ignoring systemic forces in favor of individualized framing prevents collective consciousness and social change.

For Gen-Z, whose politics are shaped within the digital sphere, the trad-wife aesthetic functions as a gauge, Some interpret it as harmless personal branding content, while others view it as "soft-core radicalization" (Proctor), and subtly normalizing conservative ideologies in mainstream content. The aesthetic therefore both reflects and shapes Gen-Z's political consciousness: it divides audiences between neoliberal tolerance of "choice" and a recognition of the systemic conservatism embedded in the content. The trad-wife, in this sense, not only projects nostalgic femininity but also brings about new forms of digital resistance, reflecting the relationship between online aesthetics and generational politics.

Soft Core Radicalization

Soft-core radicalization refers to the subtle process by which individuals are drawn toward political ideologies through seemingly apolitical or lifestyle content rather than overt extremism. In the context of the trad-wife aesthetic, this occurs through curated depictions of domesticity, motherhood, and homemaking that appear harmless or aspirational but normalize conservative gender roles as natural and desirable. Unlike overt radicalization, which relies on explicit ideological messaging, soft-core radicalization is embedded in everyday imagery: the white dresses, handmade bread, family dinners, and sunlit kitchens that Nara Smith and Hannah Neeleman present to millions of followers. As discussed in the findings, their content functions as both lifestyle branding and business, blending authenticity with marketability. This combination makes their message especially effective, because audiences encounter conservative ideals as consumable aesthetics rather than political arguments. The fact that their videos avoid overt political rhetoric further enhances their reach, since viewers interpret the content as apolitical entertainment, even while engaging with narratives of tradition, femininity, and national belonging. In this way, the trad-wife figure exemplifies soft-core radicalization, where subtle aesthetic cues and capitalist circulation reinforce ideological hierarchies under the guise of personal choice.

The political implications of this aesthetic are significant, especially when considering voter behavior among Gen-Z. As Atay and Ashlock (2022) emphasize, Gen-Z's political identities are deeply shaped by the digital environment, where online aesthetics and discourses often substitute for traditional political engagement (Atay and Ashlock). The normalization of trad-wife content, even when consumed as entertainment, can subtly shift perceptions of gender roles and national belonging. Basker et al. warns that such content functions as "soft-core radicalization," embedding conservative ideologies into seemingly benign cultural products (Basker et al.). This matters for voter turnout and political mobilization: if Gen-Z audiences interpret the trad-wife aesthetic as harmless choice feminism (Ferguson), they may disengage from structural critiques of inequality, weakening the potential for collective action. Conversely, Gen-Z creators who parody or critique trad-wife content demonstrate a growing digital resistance that can foster political consciousness in line with Nachescu's (2009) reminder that the personal is always political (Nachescu).

The trad-wife aesthetic must be situated within broader structures of power that extend beyond domestic labor and online branding. Scholars have long noted that femininity, when framed as tradition, is often mobilized to support racialized projects. As Zahay (2022) argues, the performance of femininity frequently functions as a mask for political ideology, embedding conservative values within seemingly apolitical lifestyles (Zahay). This is particularly evident in how the
trad-wife aesthetic idealizes domesticity as inherently white, middle-class, and heteronormative. Noting that there are queer and non-white smaller sub-groups of influencers that produce

‘traditional’ content. Moreover, the content tends to appreciate the practice of homesteading, rather than be charged with conservative rhetoric. Such representations echo colonial logics that cast white women as moral guardians of the nation, while simultaneously erasing or subordinating women of color whose labor historically sustained these domestic ideals (Glenn). By circulating nostalgic images of domesticity, trad-wife content reinforces racialized hierarchies of belonging, where whiteness is central to the performance of family, femininity, and national identity. The alignment of the trad-wife aesthetic with white supremacist currents also reflects the longstanding connection between gender conservatism and racial politics. As Berlet (2004) shows, women have historically played key roles in far-right and white supremacist movements, often through the performance of domesticity and motherhood as symbols of racial purity (Berlet). In digital form, the trad-wife aesthetic performs a similar function by romanticizing family, home, and reproduction as cultural values to be preserved. This process mirrors what Sykes and Hopner (2023) describe as the commodification of conservative values, where lifestyle branding transforms domestic life into a consumable product that normalizes racialized and gendered hierarchies (Sykes and Hopner). While not every trad-wife creator explicitly aligns with white nationalism, the aesthetic’s reliance on nostalgia for an imagined past aligns with exclusionary visions of the United States as a nation.

Conclusion

Ultimately, the trad-wife aesthetic is not just a personal lifestyle or branding strategy but a site where issues of gender, race, nation, and politics converge. By analyzing both the influencers and the reactionary content, this research reveals how social media operates as a political sphere where aesthetics, consumerism, and ideology intertwine. Trad wife as a digital aesthetic romanticizes domesticity through well-lit angles, dainty outfits, and picturesque familial dynamics. Subsequently commodifying the lifestyle through the marketability of its aesthetic appeal to viewers. Nara Smith and Hannah Neilman, while producing innocuous and idyllic content from a viewer standpoint, are players in shaping young voters' political identities through subtle ideological cues. It reproduces colonial and white supremacist ideals by centering whiteness and heteronormativity. In this sense, the trad-wife figure represents not merely a nostalgic fantasy but a digital performance that influences the trajectory of generational politics in the United States.

Works Cited

"A Political and Cultural Glimpse Into America's Future: Generation Z's Views on Generational Change and the Challenges and Opportunities Ahead." *PRRI*, 22 Jan. 2024, https://prri.org/research/generation-zs-views-on-generational-change-and-the-challenges-and-opportunities-ahead-a-political-and-cultural-glimpse-into-americas-future/.

Aizin, Rebecca. "Who Is Nara Smith? All About the TikTok Creator Married to Lucky Blue Smith." *People.Com*, People, 12 Apr. 2024, https://people.com/who-is-nara-smith-lucky-blue-smith-wife-8604660.

Atay, Ahmet, and Mary Z. Ashlock. *Social Media, Technology, and New Generations: Digital Millennial Generation and Generation Z*. Bloomsbury Academic, 2022.

Babb, Ellen. "Women and War: St. Petersburg Women During World War II." *Florida Historical Quarterly*, vol. 73, no. 1, Mar. 2022. *COinS*, https://stars.library.ucf.edu/fhq/vol73/iss1/6.

Basker, Sanjana, et al. "Unpacking (Alt-Right) Postfeminist Sensibilities: Theorizing the Tradwife Influencer Movement Through Case Study Analysis." *Psychology of Women Quarterly*, May 2026, p. 03616843261442447. *SAGE Journals*, https://doi.org/10.1177/03616843261442447.

Berlet, Chip. "Inside Organized Racism: Women in the Hate Movement." *American Anthropologist*, vol. 106, no. 1, 2004, pp. 180–81. *Wiley Online Library*, https://doi.org/10.1525/aa.2004.106.1.180.

Butler, Melissa A. "Early Liberal Roots of Feminism: John Locke and the Attack on Patriarchy." *American Political Science Review*, vol. 72, no. 1, Mar. 1978, pp. 135–50. *Semantic Scholar*, https://doi.org/10.2307/1953604.

Ferguson, Kathy E. "Feminist Theory Today." *Annual Review of Political Science*, vol. 20, no. Volume 20, 2017, May 2017, pp. 269–86. *www.annualreviews.org*, https://doi.org/10.1146/annurev-polisci-052715-111648.

Glenn, Evelyn Nakano. "From Servitude to Service Work: Historical Continuities in the Racial Division of Paid Reproductive Labor." *Signs: Journal of Women in Culture and Society*, vol. 18, no. 1, Oct. 1992, pp. 1–43. *Semantic Scholar*, https://doi.org/10.1086/494777.

Keegan, Beamish J. *AGE OF EMPOWERMENT: A FEMINIST DISCOURSE SUPPORTING TRADITIONAL WIVES*. 2024, https://soar.suny.edu/entities/publication/21222555-9bbd-48e2-b531-f7d8b909264d.

Nachescu, Voichita. "Radical Feminism and the Nation: History and Space in the Political Imagination of Second-Wave Feminism." *Journal for the Study of Radicalism*, vol. 3, no. 1, 2009, pp. 29–59. *Semantic Scholar*, https://doi.org/10.1353/jsr.0.0012.

Paz, Christian. "There Are Two Gen Zs." *Vox*, 27 May 2025, https://www.vox.com/politics/414266/republican-conservative-two-gen-z-young-voter-trumpier-progressive.

Proctor, Devin. "The #Tradwife Persona and the Rise of Radicalized Domesticity." *Persona*

Studies, vol. 8, no. 2, 2022, pp. 7–26. *ojs.deakin.edu.au*, https://doi.org/10.21153/psj2022vol8no2art1645.

Rogan, Frances, and Shelley Budgeon. "The Personal Is Political: Assessing Feminist Fundamentals in the Digital Age." *Social Sciences*, vol. 7, no. 8, Aug. 2018, p. 132. *www.mdpi.com*, https://doi.org/10.3390/socsci7080132.

Rogers-Cooper, Justin. "Rethinking Cold War Culture: Gender, Domesticity, and Labor on the Global Home Front." *International Labor and Working-Class History*, edited by Kate Brown et al., no. 87, 2015, pp. 235–49.

Sykes, Sophia, and Veronica Hopner. "Tradwives: Right-Wing Social Media Influencers." *Journal of Contemporary Ethnography*, vol. 53, no. 4, Aug. 2024, pp. 453–87. *SAGE Journals*, https://doi.org/10.1177/08912416241246273.

Takeuchi, Michiko. "At the Crossroads of Equality versus Protection: American Occupationnaire Women and Socialist Feminism in US Occupied Japan, 1945–1952." *Frontiers: A Journal of Women Studies*, vol. 38, no. 2, 2017, pp. 114–47. *JSTOR*, https://doi.org/10.5250/fronjwomestud.38.2.0114.
U.S.-Japan Women's Journal, no. 50, 2016, pp. 3–28.

Zahay, Megan L. "What 'Real' Women Want: Alt-Right Femininity Vlogs as an Anti-Feminist Populist Aesthetic." *Media and Communication*, vol. 10, no. 4, Nov. 2022, pp. 170–79. *www.cogitatiopress.com*, https://doi.org/10.17645/mac.v10i4.5726.

How do Marine Organisms Inspire Sustainable Product Design? By Aarohi Godge

Abstract

Modern construction and manufacturing industries contribute significantly to environmental degradation, resulting in substantial material waste, resource depletion, and carbon emissions. While traditional designs prioritize convenience over long-term ecological balance and sustainability, natural ecosystems operate through circular, waste-free systems refined over millions of years of evolution. This study investigates the structures and mechanisms of marine organisms that have evolved over millions of years, enabling them to operate in a system of resource efficiency and sustainability. This study explores how biomimicry can be integrated into product design to adapt ecological systems to human life and engineering. Systemic exploration of different marine species and organisms reveals key functional adaptations applicable across three main engineering domains: fluid dynamics (friction- reducing hydrogel coatings inspired by tuna and pilot whale), advanced adhesion and structural mechanics (non- toxic wet adhesives modeled on mussel byssal threads and impact resistant geometries from glass sponges and abalone nacre) and dynamic, responsive systems (variable-stiffness polymers inspired by sea cucumbers and fluidic solar-modulating panels modeled on squid chromatophores). Adapting these biological mechanisms enables a paradigm shift from subtractive, energy-intensive fabrication to regenerative, low-energy designs. Integrating biomimetic principles allows product designers to significantly eliminate hazardous chemicals, reduce fuel emissions, and optimize material efficiency. However, widespread industrial adoption of these biological mechanisms is constrained by the limited scalability of nanoscale manufacturing, high upfront R&D costs, and the lack of self-healing properties in static materials.

Keywords: Biomimicry, marine organisms, sustainable product design, regenerative manufacturing and eco-design

Introduction

Human manufacturing technologies and industries are growing rapidly, increasing the demand for innovative, efficient, and sustainable materials. The construction industry, in particular, depends on advanced manufacturing to develop durable infrastructure while reducing environmental impact (Carcassi et al., 2024). These practices result in substantial material waste and profound environmental damage, accelerating resource depletion, increasing carbon emissions, and accumulating industrial waste that places intense pressure on ecosystems (Korhonen et al., 2018). Amid rapid global change and technological advancements, products are being invented every day, trying to make human lives easier. However, convenience is prioritized over sustainability, contributing to adverse effects on natural ecosystems. The construction industry is responsible for 40% of natural resources consumption, 40% of waste generation, and

33% of harmful gas emissions (Van Stijn & Gruis, 2020). This underscores the critical need to adopt intensive industrial sustainability practices (Silva et al., 2024). Natural ecosystems function without producing harmful waste and run in opposition to the artificial industrial system. For instance, Nacre (a biological material found in shells) is a microscopic structure that mixes rigid minerals with soft proteins. This biological arrangement results in a fracture toughness roughly 3,000 times higher than that of the raw mineral it is made from (Zhong et al. 2023).

The marine ecosystem comprises unique, dynamic fluid challenges that have led to the development of properties by aquatic organisms to adapt and work around them. Over a billion years, marine life has evolved, resulting in the development of highly optimized morphological forms, specialized surface textures, methods and tools they use in their daily lives to maximize survival. For example, streamlined body geometries observed in fast-swimming species such as dolphins and tuna, or microscopic skin textures such as shark skin denticles, drastically reduce fluid drag and optimize energy efficiency during locomotion (Masud et al., 2025). In contrast, human-made ships and submarines waste a lot of fuel and energy due to their flat surfaces, which create too much friction in the water. While these new creations have short-term mechanical and economic efficiency, they lack a sustainable framework, leading to long-term ecological degradation. There needs to be a paradigm shift from traditional manufacturing models that cause ecological harm to one that integrates biomimicry, enabling designers to treat nature as an active blueprint and to utilize structural optimization and closed-loop lifecycles to create inherently resource-efficient products. Thus, investigating these aquatic evolutionary adaptations of marine animals would offer a direct, much-needed pathway for engineering and designing more efficient and sustainable products. By learning from nature's evolutionary solutions, designers can create products that optimize performance, minimize waste, and contribute to more regenerative manufacturing practices.

Product design can adapt biological principles to achieve sustainability goals, as natural organisms have evolved efficient, resource-conscious, and regenerative solutions over millions of years. By emulating these strategies, designers can reduce material and energy consumption, minimize waste, and create products with lower environmental damage and impact. Nature uses the same resources to sustain ecological balance without harming the environment. In contrast, many human activities exploit these resources, resulting in environmental degradation. There is a need to learn from nature by adapting processes, strategies, and structures into the design of everyday products. This can promote sustainability while maintaining functionality and shifting the focus from convenience to long-term environmental protection. Biomimicry provides a framework for attaining inspiration from nature's solutions. Hence, the present study aims to examine efficient and functional structures found in marine organisms and explore the potential application in the design of more sustainable products. The study aims to examine which marine organisms provide useful design principles, how these principles are translated into products and structures, and how biomimicry contributes to making designs more sustainable.

Literature review

Environmental challenges have intensified over recent decades, reaching unprecedented levels in the twenty-first century. Human activities have placed considerable pressure on natural ecosystems, driven by the assumption that natural resources exist primarily to meet human needs. However, the environmental consequences of this perspective have increasingly challenged this anthropocentric view, highlighting the growing disconnect between human development and the natural world. While technological advancement has transformed modern society, many innovations continue to contribute to pollution and resource depletion rather than offering sustainable solutions. For example, the production of cement and steel for construction releases large amounts of carbon dioxide, while single-use plastics manufactured for convenience contribute to widespread land and marine pollution (Nations, 2019). Consequently, researchers and designers have increasingly turned to nature as a source of inspiration, giving rise to the field of biomimicry.

Biomimicry is an emerging interdisciplinary approach that addresses complex design challenges by emulating nature's strategies, structures, and processes. The term was introduced by Janine M. Benyus in 1997 and is derived from the Greek words **bios** (life) and **mimesis** (imitation). Related concepts, including biomimetics, bio-inspired design, and biologically inspired engineering, are used to describe innovations derived from natural systems (DesignLens: Life's Principles, n.d.). Over approximately 3.8 billion years of evolution, nature has developed highly efficient and resilient solutions that frequently outperform human-engineered systems while using fewer resources and generating minimal waste. Natural systems operate through circular resource use, energy efficiency, self-organization, and adaptability, making them valuable models for sustainable innovation. By learning from these biological principles, biomimicry offers a framework for developing products, technologies, and systems that not only enhance functionality but also promote environmental sustainability and support ecological restoration. This review examines the relationship between biomimicry and sustainable design, with particular emphasis on the adaptations of marine organisms (DesignLens: Life's Principles, n.d). Biomimicry's "Life's Principles" include adapting to changing conditions, being locally attuned and responsive, using life-friendly chemistry, being resource-efficient, integrating development with growth, and evolving to survive (Figure 1; DesignLens: Life's Principles, n.d.).

Figure 1: DesignLens: Life's Principles

Natural systems adapt to different environments over time through evolution and by responding to new conditions, helping them remain resilient and able to survive. The systems work well with their local environment and are built and function according to the specific conditions of their surroundings. Nature uses safe chemicals, avoiding harmful substances and processes. Organisms use resources wisely, making the most of their energy and materials while keeping waste to a minimum. Another principle is that growth and development occur in ways that do not consume more than what is available. Finally, all systems can change and improve over time. Together, these principles provide a framework for product designers to create regenerative, resilient, and sustainable solutions that utilize natural processes rather than opposing them and causing damage to nature (Blasiak et al., 2022).

This systems-based approach is a direct consequence of applying Life's Principles. Biomimicry teaches designers to think in systems rather than focusing solely on the individual performance of each product. It encourages integrating a product into a larger ecological or industrial system to ensure regeneration, resilience, and feedback loops. In natural ecosystems, materials are continuously cycled so that every output becomes a resource for another organism or process, resulting in little to no material waste. It is the ultimate pathway to sustainable design (Blasiak et al., 2022).

While traditional sustainable design often focuses on developing technologies that use less energy and increase recycling, biomimicry goes even further by studying ecosystems and learning from how these creatures build entirely regenerative systems where little to no waste is produced. An example of sustainable design through biomimicry is the tubercles on humpback whale flippers (see the "Movement" section). They have inspired the development of wind turbine blades, which are simultaneously "bio-inspired" (i.e., the structure of a natural material inspires them), "biomorphic" (i.e., they resemble the shape of a living thing), and "biomimetic" (i.e., they mimic the structure of the flipper). They are also an example of "biomimicry" (i.e.,

they did not directly inspire the wind turbine, but rather a more sustainable, optimized blade design; Blasiak et al., 2022).

Shark skin is covered in very small, ribbed features that impede water flow, making swimming more efficient, and help to prevent microbial adhesion (Dundar Arisoy et al., 2018). Mussels can adhere to a variety of surfaces, including wet, organic, rough, and water-conditioned environments, using long, thin protein threads (Claverie et al., 2020). Octopuses demonstrate flexibility in their arm structure due to the use of suction cups that allow them to latch onto and remove from surfaces with great precision and manual dexterity (Tramacere et al., 2015; octopus suction biomechanics). Each of these organisms exhibits physical adaptations that facilitate their movement, protection from predators, and adherence to a variety of surfaces. Therefore, they inspire the design of materials better suited than current designs to facilitate marine animal movement, protection, and adhesion (Li et al., 2025).

Discussion

We will now discuss and analyze the fluidic, structural, and adaptive capabilities of various marine creatures. We will read in detail how biological principles can be translated into industrial engineering, evaluating sustainable profiles and the key opportunities and bottlenecks encountered in product design.

Fluid Dynamics and Surface Topology - topological fluid dynamics is a young mathematical discipline that studies the topological features of flows with complicated trajectories and their applications to fluid motions, group-theoretic and geometric points of view on various problems 2026). Typical, traditional human designs oppose fluid resistance with highly polished, smooth surfaces. However, marine organisms such as tuna and pilot whales demonstrate that microtextured hydrodynamic boundary layers are far more efficient at managing turbulent drag.

Tuna has a highly streamlined body paired with a specialized outer skin that creates micro-turbulence at the air-water boundary, drastically reducing fluid resistance at higher speeds. Tuna's structure is applied to commercial shipping containers through specialized hull coatings. Mimicking the hydrogel properties of tuna skin creates a microscopic layer of trapped water on the hull, minimizing friction with the surrounding ocean (marine coatings | nippon paint holdings co., ltd., 2026). This is commercially validated by Nippon Paint Marine's LF-Sea series, which uses biomimetic hydrogel water-trapping technology to reduce drag and fuel consumption (sustainability - nippon paint marine, 2023).

Dolphins (Pilot whale) possess a dynamic, self-cleaning skin surface that naturally sheds fouling organisms such as algae and barnacles while actively dampening drag forces. Bionic ship propeller coatings and marine turbine blades can replicate this self-cleaning, low-drag skin structure, thereby enabling maritime machinery to operate at peak performance without relying on toxic heavy-metal coatings. Described in the overview study by B. Bhushan, published in the

Royal Society A Journal, which analyzes the microscale structures of marine mammal skin. (Bhushan, 2009).

Whale (Humpback) has bumpy ridges called tubercles along the leading edge of its flippers, which channel fluid and alter local flow to prevent aerodynamic stalling at high angles of attack. It can be replicated in the design of aerodynamic fan blades, industrial ventilation systems, and wind turbines to generate significantly more power with reduced operational acoustic noise. This is analyzed in the functional biological mechanics framework by (Bhushan, 2009)**.**

Manta rays use continuous undulating body waves to channel kinetic energy, moving smoothly through strong ocean currents with minimal energy loss. This characteristic translates into hydrokinetic energy membranes that wave like flags in water currents to generate clean electricity without requiring high-speed spinning blades. Archived in the engineering documentation for the EEL Energy Tidal Generator at the U.S. Department of Energy OpenEI Marine Energy Database (Energy Information and Data | OpenEI.org, n.d.).

Advanced Adhesion and Materials Mechanics - Advanced Adhesion and Materials Mechanics is an interdisciplinary field that studies the physical, thermodynamic, and mechanical principles governing bond formation, interfacial load transfer, and failure dynamics between joined materials under mechanical or environmental loads (Turusov, 2019; Van der Sluis et al., 2018). Adhering to wet surfaces or building lightweight, crush-resistant structural geometries are two of the most complex challenges in materials science. Marine organisms solve these with elegant chemical and mechanical adaptations.

Mussels: Mussels produce byssal threads made of unique proteins that undergo rapid enzymatic cross-linking. This creates a really strong, water-resistant bond that adheres to wet, slick, irregular surfaces under turbulent wave conditions. Replicating this chemistry enables the formulation of formaldehyde-free, non-toxic plywood adhesives and biocompatible medical glues that seal surgical incisions in wet physiological environments (Biomimicry Institute, 2016).

Remora Fish (suckerfish) uses a modified dorsal fin that functions as a highly specialized suction disc, with rows of friction-generating lamellae to anchor itself to fast-moving hosts reversibly and without tissue damage. This mechanical design is utilized to engineer reversible underwater robotic hitches. These grippers allow unmanned underwater vehicles (UUVs) to hitchhike on commercial vessels, conserving battery power during extended surveys, documented in Science Robotics (Wang et al., 2017).

Glass Sponge (Venus' Flower Basket) possesses a highly structured, hierarchical silica lattice skeleton. This unique geometric arrangement perfectly distributes localized stresses, preventing structural collapse under crushing ocean pressures. This adaptation of lattice architecture is used

in aerospace frames, skyscraper designs, and bridge structures to optimize load-bearing capacity while minimizing material weight. Published by the American Physical Society (Fernandes et al., 2021).

Abalone (Mollusk) builds a shell lined with nacre (mother-of-pearl), structured like a dense, microscopic brick-and-mortar wall, which diffuses energy, making it highly resistant to fracture. Microfracture-resistant composite glass and protective ceramic structures are stronger than regular glass and ideal for aerospace windows and protective casings. Primary academic study on bio-inspired tough ceramics published in *Nature Materials* (Bouville et al., 2014). This study explores how natural **nacre** (mother-of-pearl)—found in the inner layer of abalone shells—achieves extraordinary toughness despite being composed of over 95% inherently brittle mineral (calcium carbonate) held together by a tiny fraction (5%) of organic biopolymers.

Boxfish possesses a lightweight, hexagonal bony shell (cerapace) that absorbs and distributes high-impact forces—a feature used in bionic car frame architectures to increase passenger safety while significantly reducing vehicle mass. Documented in mechanical structure studies (Bhushan, 2009).

Sea urchin possesses a five-part jaw mechanism (Aristotle's lantern) with self- sharpening teeth capable of grinding through solid coastal rocks. Replicated by space exploration engineers to design highly efficient mechanical grabbers and sediment samplers to collect soil on Mars without needing heavy machinery. Analyzed in the mechanical extraction strategy guide on (AskNature - Innovation Inspired by Nature, 2016) and NASA Tech Briefs.

Gecko employs millions of foot hairs, called setae, that cling to surfaces using Van der Waals forces. Applied in glue-free, heavy-duty industrial tapes and robotic grippers that leave no sticky residue or chemical contaminants. This application is analyzed in the dry adhesion materials research by Bhushan (2009).

Dynamic and Responsive Systems - Dynamic and Responsive Systems (often referred to as smart or adaptive materials systems) are integrated physical or material architectures capable of sensing environmental stimuli—such as temperature, stress, light, or electric fields—and dynamically altering their properties or behaviors predictably and reversibly (Walther, 2019). Human-made materials are typically static. In contrast, marine organisms use responsive, stimulus-sensitive materials to adapt to their surroundings in real time.

Sea cucumbers can alter the cross-linking behavior of their collagen-based dermis, shifting their skin from highly pliable to rigid and tough within seconds of detecting environmental stress. This mechanism is used to develop adaptive neurological implants and brain microelectrodes.

The implants remain sufficiently stiff for precise surgical insertion but quickly soften upon contact with physiological moisture, minimizing brain tissue scarring (Zhang et al., 2017).

Caribbean reef squid possess specialized muscles that expand and contract color-changing pigment sacs, called chromatophores, in their skin to regulate light reflection and solar absorption dynamically. Replicated in smart windows and dynamic building panels that shift fluids through microchannels to alter window tint, cutting heating and cooling costs by up to 43%. Architectural engineering study on fluidic systems for solar modulation by the University of Toronto Engineering (Kay et al.).

Salp (Gelatinous Zooplankton) glides through the ocean via a highly efficient form of jet propulsion, drawing water into its barrel-shaped body and contracting muscles to pump the water out. Applied to micro-UUV propulsion systems and soft-robotic pumps to navigate delicate coral reefs silently without generating wake turbulence or physical damage.Hydrodynamic jet propulsion analysis published in the Journal of Fluid Mechanics (Worster, 2019).

Clinging jellyfish secrete a specialized underwater mucus mesh that precisely captures microscopic nutrients and particles suspended in moving water currents and translates into biomimetic ocean filters designed to screen microplastics and other suspended particles out of municipal wastewater lines without clogging like traditional mesh filters. Environmental engineering research on filter-feeders published in ACS Environmental Science. (Anon, 2019). The study examines how jellyfish (*Aurelia* and other medusae) produce sticky, protein- and mucin-rich bio-gels (mucus) as a defense or feeding mechanism.

Pantisocratic Coral (and Hard Corals) extracts ambient calcium carbonate from seawater to mineralize dense, durable skeletons, locking carbon away permanently. Utilized in carbon-negative cement manufacturing (e.g., CarbiCrete), which cures concrete using trapped industrial carbon dioxide instead of heat. Innovation profile on coral-inspired carbon-sequestering building materials published on the Biomimicry Institute / AskNature Strategy Guide. (AskNature - Innovation Inspired by Nature, 2016). This research focuses on biomineralization—specifically how hard corals extract dissolved carbon dioxide and calcium ions from seawater to construct durable calcium carbonate skeletal structures (aragonite).

Transition into Design and Sustainability Benefits

Converting biological systems into commercial applications requires going beyond surface-level characteristics to replicate fundamental organic processes. A key advantage lies in energy conservation and emissions reduction: using coatings inspired by the biology of tuna and dolphins on commercial vessels effectively minimizes hull drag. A 15% improvement in hydrodynamics delivers an approximate 10% decrease in fuel consumption, substantially dropping total greenhouse gases from maritime shippinghttps (Nippon Paint Marine, 2022).

Additionally, biomimicry targets chemical safety through toxicity mitigation: adhesives modeled on mussel chemistry deliver high-strength bonding performance without releasing hazardous volatile organic compounds or formaldehyde, safeguarding the health of manufacturing staff and ensuring safer home environments for end users (Openei.org, 2022). Finally, clean power development is enhanced through renewable electricity systems: replicating the ongoing, wavy swimming motions of the Manta Ray and Eel enables the creation of hydrokinetic membranes that oscillate naturally within aquatic currents, generating clean electrical energy without the environmental hazards of high-speed mechanical blades (Openei.org, 2022).

These findings and research demand a profound shift in how product designers select, structure, and manufacture materials and products. In the past, industrial design has relied heavily on subtractive fabrication- shaping a product by cutting away material from a larger standardized block. The functional design paradigm forces designers to shift from purely aesthetic roles to multidisciplinary coordination. By using databases like AskNature and coordinating with marine biologists, engineers can transition from artificial chemistry to organic, low-energy, and regenerative designs.

Despite the benefits, critical engineering and economic bottlenecks prevent bio-inspired concepts from scaling. Scalability Barrier: Current high-throughput manufacturing, such as injection molding, cannot cheaply or easily replicate nanoscale biological architectures, such as dolphin skin or gecko foot hairs (Bhushan, 2009). Next, the Static vs. Living Paradox: Unlike dynamic living tissue, synthetic biomimetic surfaces lack cellular self-healing properties. When scratched by debris, these static coatings degrade rapidly in harsh marine environments (Bhushan, 2009). Lastly, economic disadvantages: Re-engineering traditional manufacturing pipelines requires substantial upfront R&D investment, making it difficult to replace cheap, established petroleum-based polymers or to replicate complex mechanisms such as the sea urchin's jaw (AskNature - Innovation Inspired by Nature, 2016).

Conclusion

This exploratory study highlights the significant potential of biomimicry in the transition toward sustainable product design and regenerative manufacturing. By analyzing the structural, fluidic, and adaptive mechanisms of diverse marine organisms—ranging from the drag-reducing skin of tuna and dolphins to the dynamic, tissue-softening collagen of sea cucumbers—this paper demonstrates how biological adaptations can be successfully and strategically translated into low-energy, non-toxic, and high-efficiency sustainable engineering solutions. Despite current technical and economic limitations, including the challenges of scaling nanoscale fabrication, high initial research and development costs, and the lack of self-healing capabilities in static synthetic materials, these findings underscore the transformative impact of nature-inspired design. Future research should focus on improving fast 3D printing to replicate complex biological structures at scale, while also developing self-repairing materials that enhance product durability and efficiency. Additionally, future studies should investigate standardized life-cycle

assessment frameworks specifically tailored for biomimetic products to measure their long-term environmental benefits and accelerate industrial adoption.

Works Cited

AskNature - Innovation Inspired by Nature. (2016). [online] AskNature. Available at: https://asknature.org [Accessed 7 Aug. 2026].

Bhushan, B. (2009). Biomimetics: lessons from nature–an overview. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1893), pp.1445–1486. doi:10.1098/rsta.2009.0011.

Biomimicry Institute. (2016). *PureBond® Technology: Wood Glue Without Formaldehyde*. Biomimicry Toolbox.

Blasiak, R., et al. (2022). A Forgotten Element of the Blue Economy: Marine Biomimetics and Inspiration from the Deep Sea. *PNAS Nexus*, 1(4), p.pgac196.

Carcassi, O. B., Salierno, R., Falcinelli, P. A., Paoletti, I. M., & Ben-Alon, L. (2024). Upscaling natural materials in construction: Earthen, fast-growing, and living materials. *Sustainability*, 16(18), 7926.

Claverie, M., McReynolds, C., Petitpas, A., Thomas, M. and Fernandes, S.C.M. (2020). Marine-Derived Polymeric Materials and Biomimetics: An Overview. *Polymers*, 12(5), p.1002. doi:10.3390/polym12051002.

Dundar Arisoy, F., Kolewe, K.W., Homyak, B., Kurtz, I.S., Schiffman, J.D. and Watkins, J.J. (2018). Bioinspired Photocatalytic Shark-Skin Surfaces with Antibacterial and Antifouling Activity via Nanoimprint Lithography. *ACS Applied Materials & Interfaces*, 10(23), pp.20055–20063. doi:10.1021/acsami.8b05066.

Energy Information and Data | OpenEI.org. (n.d.). [online] openei.org. Available at: https://openei.org [Accessed 7 Aug. 2026].

Journal of Fluid Mechanics | Cambridge Core. (2019). [online] Cambridge Core. Available at: https://www.cambridge.org/core/journals/journal-of-fluid-mechanics [Accessed 7 Aug. 2026].

Korhonen, J., Honkasalo, A., & Seppälä, J. (2018). Circular Economy as an Essentially Contested Concept. *Journal of Cleaner Production*, 175, pp.544–552.

Li, L., Qin, B., Gao, W., Li, Y., Zhang, Y., Wang, B., Kong, S., Wang, J., He, D. and Yu, J. (2025). Bioinspired underwater soft robots: from biology to robotics and back. [online] arXiv.org. Available at: https://arxiv.org/abs/2508.11883 [Accessed 7 Aug. 2026].

Marine coatings | NIPPON PAINT HOLDINGS CO., LTD. (2026). [online] NIPPON PAINT HOLDINGS CO., LTD. Available at: https://www.nipponpaint-holdings.com/en/business/marine/ [Accessed 7 Aug. 2026].

Nature (2019). Nature. [online] Nature. Available at: https://www.nature.com/ [Accessed 7 Aug. 2026].

Nippon Paint Marine. (2022). *Sustainability Marine Coatings | Biocide Free | Nippon Paint Marine*. [online] Available at: http://www.nipponpaint-marine.com/sustainability/ [Accessed 7 Aug. 2026].

Openei.org. (2022). *PRIMRE/Databases/Projects Database/Devices/EEL Energy Tidal Generator | Open Energy Information*. [online] Available at:

https://openei.org/wiki/PRIMRE/Databases/Projects_Database/Devices/EEL_Energy_Tidal_Generator [Accessed 7 Aug. 2026].

Sustainability - Nippon Paint Marine. (2023). [online] Nippon Paint Marine. Available at: https://www.nipponpaint-marine.com/sustainability/ [Accessed 7 Aug. 2026].

Turusov, R. A. (2019). Adhesion and Adhesion Mechanics. Features of the Theory and its Possibilities. *Materials Science Forum*, 974, pp.638–645.

United Nations (2019). *Plastics – fueling oil demand, climate change and pollution | United Nations*. [online] United Nations.

Wang, Y., Yang, X., Chen, Y., Wainwright, D. K., Wu, C. P., Gong, Z., Liang, M., Lu, H., Liu, Y., Kou, Z. L., Zhang, Y. F., & Wen, L. (2017). A biorobotic adhesive disc for underwater hitchhiking inspired by the remora suckerfish. *Science Robotics*, 2(10), eaan1074.

Zhang, X., Sun, L., Yuan, J., Sun, Y., Gao, Y., Zhang, L., Li, R., & Chen, M. (2017). The sea cucumber genome provides insights into event-driven tissue regeneration and body wall modification. *Science Advances*, 3(10), e1700789.

The Physics of Color Rendering and Emotional Impact: How LED Spectrum Optimization Transforms Audience Perception in Musical Theater By Hridaan Jhunjhunwala

Abstract: In any on-stage performance, especially musical theater, lighting plays a crucial role in determining the visual impact, attracting audiences' attention, and even shifting the mood. The result becomes prominent as LED replaces traditional lighting setups, with a different range of light wavelengths. Addressing its growing adoption in stage lighting, this study investigates how variations in LED lighting characteristics, specifically Color Rendering Index (CRI) and color temperature, affect audience responses during a theatrical performance. A comparative experimental approach was adopted in which participants viewed the same scene under two different LED lighting conditions, including a warm, high-CRI setup and a cooler lighting setup. Audience perceptions were measured using a questionnaire that assessed color quality, emotional engagement, atmosphere, and overall enjoyment. The findings reported that high-CRI lighting conditions received consistently higher ratings, also supported by existing literature. The results suggest that optimizing LED spectrum characteristics enhances not only the visual quality of stage presentations but also the emotional experience of viewers.

Keywords: LED Lighting ,Color Rendering Index (CRI), Color Temperature, Audience Perception, Emotional Engagement, Musical Theatre, Stage Lighting

Chapter 1: Introduction

1.1 Importance of Lighting in Musical Theatre

Lighting has been one of the fundamental aspects of theatrical performances. Similar to motion pictures, lighting is not only limited to highlighting the important characters, but also has the power to create the atmosphere for the story. It would help guide audiences' attention towards something specific, develop the mood for scenes, and support the overall storytelling process (Kashaka, 2025). Unlike motion pictures, where scenes replicate real-life surroundings, theatrical performance uses lighting differently. It uses lighting not to replicate the real-life surroundings, but to mimic the very essence of those surroundings, so different emotions can be provoked from the audience as well. As reported, if lighting can be used in an innovative way, it can go beyond illumination and extend to "tell your story, set your tone, and leave a lasting impression" (ABS, 2022).

Figure 1.1.1: Stage Lighting's Importance
Source: (Liang, 2025)

As shown in Figure 1.1, lighting serves as more than a source of illumination during theatrical performances. Through variations in lighting, such as brightness, color, and intensity, gaffers and lighting designers can influence the audience's overall experience of the live performance.

1.2 LED Technology in Contemporary Theatre

Figure 1.2.1: Modern LED lighting vs Traditional Stage Lighting
Source: (Zhang, 2025)

Over the decade, "light-emitting diode" (LED) technology has become a common integrated part of theatrical performance. Compared to its predecessors, such as "Incandescent" and "Fluorescent", it offers higher energy efficiency, longer operational time, and greater control over color and intensity (Barrett & Jeffery, 2026). Also, as seen in Figure 1.2.1, modern LED fixtures allow designers to create various lighting effects while consuming less power. This feature makes them optimal for use in both professional and educational theatrical performances.

1.3 Color Rendering and Color Temperature

LEDs have two important characteristics, which are color rendering and color temperature. Color rendering refers to the ability to reveal the colors of various objects faithfully. It is compared by using a standard reference light source, which is sunlight. This property is measured by using “Color Rendering Index” (CRI), in which a higher value indicates higher accuracy in color production. Color temperature describes the visual appearance of the light, which ranges from warm yellows to cool blues. It is measured in “Kelvin” (K) with a range of 1000K to 10000K, in which 2000K represents warm yellow tones and 5600K refers to cool blue tones, as referred to in Figure 1.3.1. These properties determine the visual quality of a theatrical performance experienced by the audience.

Figure 1.3.1: Different Colour Temperatures of LED
Source: (Upward Lighting, 2023)

Especially in musical theatre, where audiences heavily rely on visual information to interpret the setting, characters, or the emotions of the particular situation, lighting plays a vital role in controlling them (Zhang et al., 2023). Indeed, costumes, makeup, stage scenarios, and performers’ facial expressions and speech contribute significantly to the storyline process, but the effectiveness of such visual elements would equally depend on how those features are illuminated, rather than the sign. As Rossi et al. (2024) suggested, poor lighting conditions can distort colors, reduce visual clarity, and alter the appearance of performing their ads and stage objects. In contrast, a well-designed lighting stage with suitable color rendering and temperature can enhance the visual realism and support the artistic message which was intended by the performers. Hence, identifying which lighting should be used and what its quality should be is necessary in shaping both the audience’s perspective and generating their emotional response in a musical theatre.

1.4 Research Gap and Purpose of the Study

Existing studies on the psychological effects of lighting suggest that different lighting conditions can affect human emotions, attention, and perception. For instance, warm lighting instigates feelings of comfort, intimacy, and relaxation, whereas cooler tones may evoke alertness, tension, or even emotional distance (Shahidi et al., 2021). Similarly, lights with high CRI values can improve the appearance of costumes, scenery, and performers, leading to an

enhancement in audiences' immersion and engagement. However, the combined effect of LED color rendering and color temperature on audience perception during theatrical sessions is an area that will require further thorough investigation.

This study aims to speculate how variations in LED lighting can affect audiences' emotions and visual perception during a theatrical performance. It aims to compare audience responses presented towards the same scene with different LED light settings. This will help to determine which LED lighting conditions with different CRI values and color temperatures can affect emotional engagement and the perceived quality of the stage visuals.

1.5 Research Question, Objectives, and Hypothesis

To integrate the result within the findings, this paper will follow an explicit research question regarding "How do variations in LED color rendering index (CRI) and color temperature influence audience perception and emotional engagement during a musical theatre performance?"

Furthermore, there are 3 objectives created to simplify the results, which are received from the survey questionnaire that follows the research question. These are mentioned below:

- To examine the impact of different LED lighting conditions on audience perception of color quality during a musical theatre performance.
- To investigate the influence of LED color rendering and color temperature on audience emotional engagement and overall viewing experience.
- To compare audience preferences between different LED lighting conditions and identify which one they prefer the most.

It has also been speculated that LED lighting conditions with higher CRI values and warm color temperatures can produce more natural-looking colors. It can also improve the quality of the theatrical performance and build strong emotional engagement among audience members, compared to lighting with lower color-rendering quality or cooler color temperatures.

Chapter 2: Background and Literature Review

2.1 Physics of Light and Color

The perception of color begins with the physical properties of light, where the light interacts with the object and observers to present through reflection, absorption, and transmission. Light can be in different wavelengths, but, as NASA (2023) reported, human eyes can only detect "wavelengths from 380 to 700 nanometers", making them visible on the electromagnetic spectrum (Figure 2.1.1). Prangnell (2016) explained it further, stating that the human visual system interprets these wavelengths through specialized photoreceptor cells located in the retina, allowing individuals to distinguish different colors and levels of brightness. Similarly, Hurlbert & Yu (2025) disclosed human inclusion in detecting color, suggesting that color perception is not only determined by the wavelength of light but also dependent on individuals' brains' ability to interpret variations in illumination and visual context. Their

explanation exemplifies the reason why the same object can appear differently under different lighting conditions.

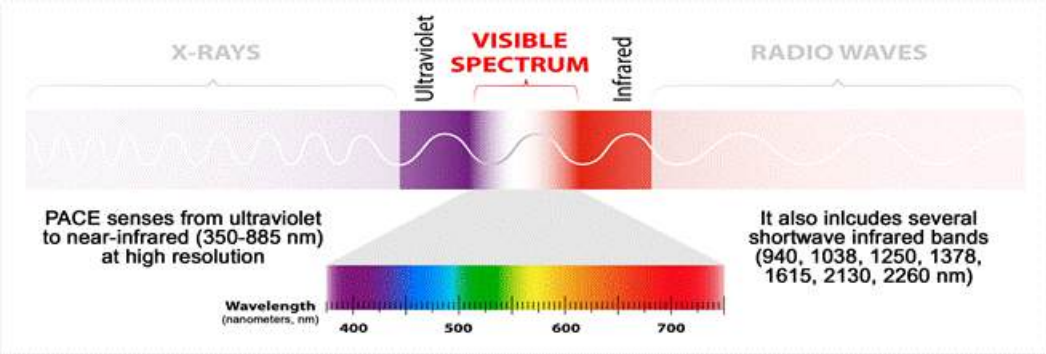


Figure 2.1.1: Visible Wavelengths to Human Eyes
Source: (NASA, 2024)

Colour is only visible when lights and the physical objects are interconnected and reflected to the viewer. However, not all of the light wavelengths are reflected, as some of them are absorbed by the surface object as well. According to Marlow et al. (2022), color plays a significant role in visual perception because it assists individuals in identifying shapes, depth, and spatial relationships within an environment. The impact of color perception beyond its decorative agenda is also mentioned by Xie et al. (2022), as they further added how different colored lighting conditions can generate distinct affective impressions, demonstrating that color perception extends beyond visual recognition and is eligible to influence emotional responses as well.

In the case of a stage performance environment, the color perception and its influence by different lighting spectra further highlight the connection between light and color. As Smith & Coleman (2021) described, theatrical lighting is a combination of artistic creativity and physical science, where the directors or lighting experts manipulate the brightness, color, and direction of the light to shape the viewers' perception from the audience seat. When performing in an arts setting, colored lighting contributes to the visual atmosphere of a production and supports storytelling by differentiating performers' characters, the mood of the storyline, different expressions, and situations. In this matter, as Basa (2023) noted, lighting design and color schemes are capable of creating visual illusions that enhance audience engagement. They further indicated that effective lighting design can significantly strengthen the artistic impact of a performance and make it look more appealing. Similarly, Zhou et al. (2025) commented that variations in lighting color can significantly affect mood and creative responses surrounding the illuminative environments, which can have a direct psychological impact on the audience.
As LED technology started to be integrated within theatres around the world, the importance of understanding the connection between light and color became increasingly important. In this matter, as Asuquo (2025) explained, modern theatre productions are progressively replacing traditional incandescent systems with LED lighting. The reason behind the shift is its energy efficiency, flexibility, and enhanced color control, which are essential for theater performance

and connecting audiences. Hence, understanding the physics of light and color is essential to understanding how modern lighting technologies like LEDs can be used to influence audience perceptions and emotional experiences.

2.2 LED Spectral Distribution

A light-emitting diode (LED) is an electronic component designed specifically to emit light by using a semiconductor when the current flows through it. LEDs are fundamentally different from their predecessors, such as incandescent and fluorescent, which emit light by heating the filament inside the bulb by passing electric current. Unlike its predecessors, Pattison et al. (2018) mentioned that LEDs can operate with greater energy efficiency, have longer lifespans, and provide enhanced control over color output. As a result, LEDs have become a crucial part of the lighting system in contemporary theatrical productions.

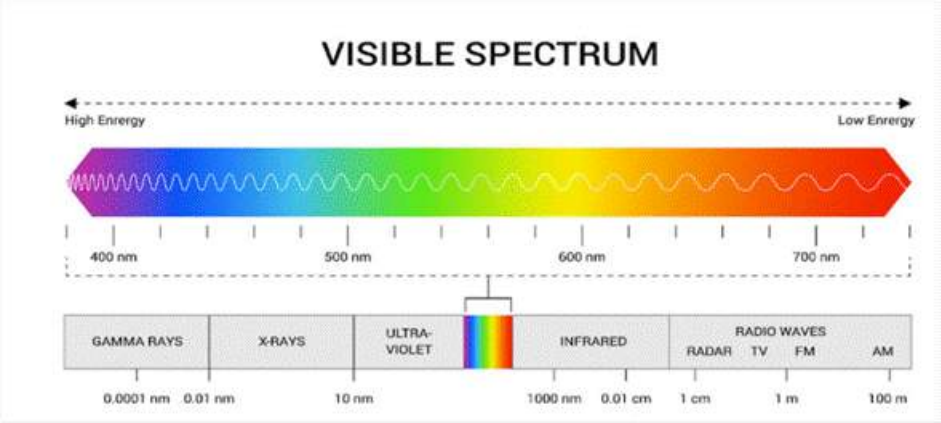


Figure 2.2.1: Visible Spectrum

Source: (Lifocolor, 2025)

A key characteristic of LED lighting is the “Spectral Power Distribution” (SPD), which describes the power output of a light source across different wavelengths of the visible spectrum, as seen in Figure 2.2.1. According to Wu et al. (2026), SPD plays a critical role in determining LEDs’ visual quality by influencing color appearance, color rendering, and overall lighting performance. As shown in Figure 2.2.2, LED and incandescent light sources exhibit different spectral distributions, which influence color rendering and visual appearance. A standard LED had a blue light bulb coated in a specific yellow powder known as phosphor. When the blue light from the bulb hits the phosphor coating, they mix and emits light that a viewer’s eye perceives as white. Moreover, LEDs can generate light through specific wavelength peaks, and the light’s distribution depends on the design of the LED and the phosphor materials used in its development.

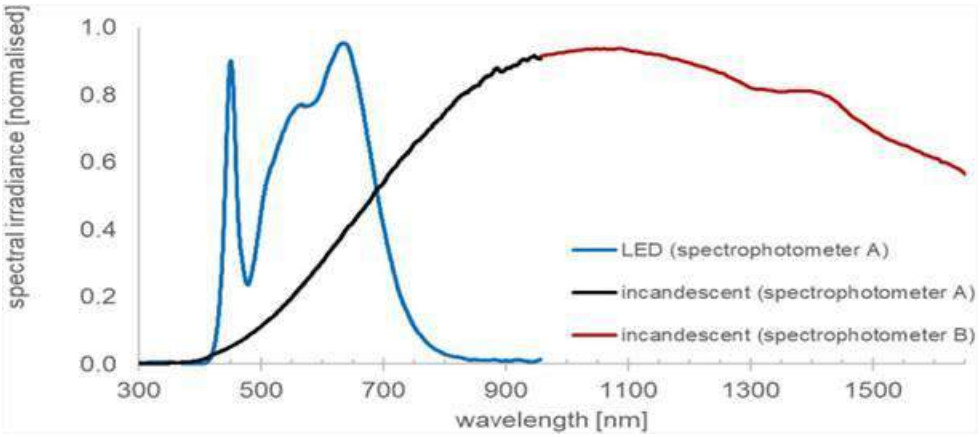


Figure 2.2.2: Comparison of Spectral Power Distribution between LED and Incandescent Light Sources

Source: (Barrett & Jeffery, 2026)

Research demonstrates that variations in LEDs' spectral distribution can significantly influence how objects are perceived under illumination. According to Lee & Yoon (2021), LEDs with optimized SPD can provide improved color rendition and increase the visual appeal of the environments. Similarly, Alrajhi et al. (2023) demonstrated that LEDs with distinct spectral compositions can produce distinct visual effects, which highlights the importance of how much energy a light source emits at every individual color rather than simply increasing the overall brightness of the light.

Furthermore, spectral characteristics also become important within theatrical environments because audiences constantly perceive the visual information of costumes, performers' expressions, and stage elements. Chan & Cheung (2022) argued that advancements in LED technology have fundamentally transformed contemporary theatre by allowing room for more creative expression through visual lighting and effects. This view was also supported by Hossain & Yadav (2025), who further stated that LED installation in art venues offers more flexibility in creating dynamic effects and provides greater control over color than traditional systems. Furthermore, Oliveira et al. (2024) found that well-designed stage lighting can contribute positively to providing visual comfort, which is an important factor for audiences' emotional engagement during theatrical performances.

Therefore, understanding LED spectral distribution is essential to evaluate color rendering and audience perception in theatres. The spectral composition of the LED determines not only how colors appear on stage but also how effectively it allows the performance to communicate to its audience by creating the intended visual and emotional atmosphere.

2.3 Color Rendering Index (CRI)

"Color Rendering Index" (CRI) refers to the quantitative scale used to compare the accuracy between science-based results and natural visualization under sunlight when a light source reveals the color of an object. It is one of the most important measures, ranging from

scale 0 to 100, used to evaluate the quality of a light source, as shown in Figure 2.3.1. According to Oh & Lim (2023), lighting systems with high CRI values more closely replicate the characteristics of natural daylight, allowing objects and surfaces to appear more realistic. Siniscalco (2022) also highlighted a similar response as he argued that color rendering is a critical consideration in lighting design because the perceived appearance of an object not only depends on the object's surface but how it is illuminated by the lighting's spectral properties.

Figure 2.3.1: High vs Low CRI
Source: (Take Three Lighting, 2026)

Especially in performance environments like stage setup or theatre, color rendering is critical to have a direct influence on audiences' perception. Lee & Yoon (2021) found that LED lighting systems with improved color rendition enhance the visual quality of interior spaces and contribute to more favourable viewer experiences. Especially since in a theater setting, suddenly interrupt visual information from different elements such as costumes, scenery, makeup, and performers. Improving color rendition becomes particularly important. Following this statement, Sokari (2025) emphasized that design elements, including lighting quality, significantly affect audience engagement and overall theatrical experience. Bell (2013) also highlights the strong influence of color on audiences' response while discussing "theatre science design", suggesting that accurate color presentation is essential for effective visual storytelling.

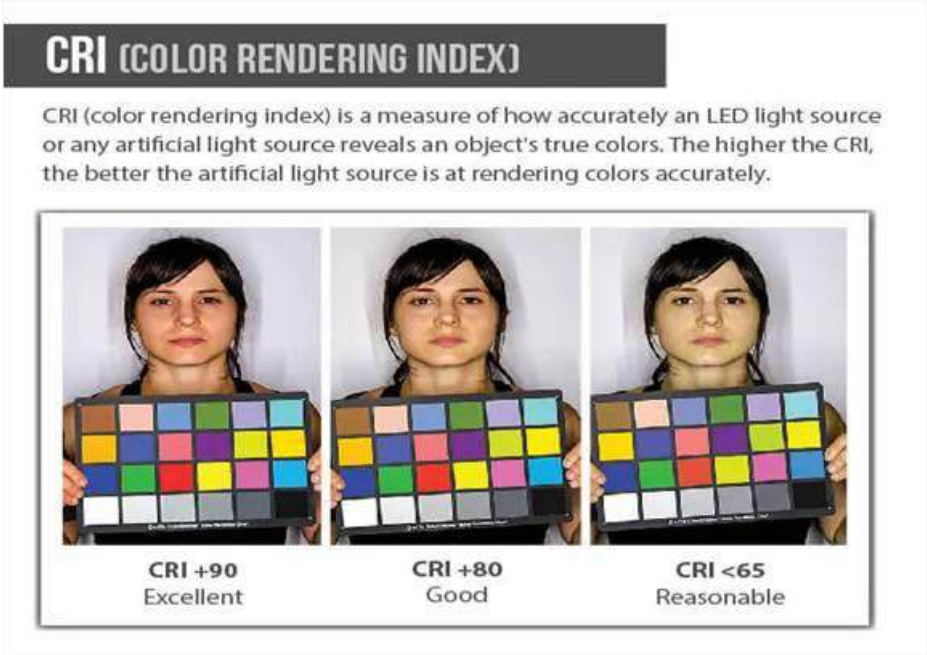


Figure 2.3.2: High-CRI Determine the Skin Colour Best
Source: (Led Lucky, 2022)

A most significant application of high-CRI lighting can be determined when reproducing human skin tone (Figure 2.3.2). Many researchers have supported this statement by highlighting light's influence on performers' visual appearance. A study by Cao et al. (2020) demonstrated that viewers generally prefer lighting conditions that reproduce skin tones naturally and consistently across different ethnic groups. Also, Hanlon et al. (2022) explained that variations in spectral power distribution can alter color consistency and visual appearance, especially where accurate skin representation is important. Even beyond the aesthetic consideration, color rendering equally impacts visual adaptation and the comfort of the visual representations. For example, Zhao et al. (2020) found that differences in LED color rendering can influence how the human eye adapts to changing lighting environments. Bishop & Chase (2023) similarly argue that effective lighting design should consider visual perception rather than relying solely on illuminance measurements because it affects the mood of the environment as well.

Researchers have also recently focused on the broader psychological effects of lighting quality to determine both the quality of the visually effective environments and the audience's perception. Supporting this, Zhou et al. (2025) directly reported that lighting color and illuminance can influence mood and creative performance. The same views are highlighted by Sarinastiti et al. (2024), who showed how the interaction between lighting and visual design elements plays a crucial role in shaping audience experiences during musical performances. All their studies and reports highlight the importance of high-CRI lighting in not only improving color accuracy but also enhancing overall audience engagement and emotional connection during the performance.

2.4 Color Temperature and Emotional Response

Color temperature refers to the visual appearance of the light and is measured by using the Kelvin (K) scale. Colors with lower temperatures around 2000-3000K produce a warm orange-yellow tone and gradually shift to bluish tones around 5000K. It is a physical property of light, and it can also influence human emotions, mood, and perception, which makes it an important consideration in theatrical lighting design, as seen in Figure 2.4.1.

Figure 2.4.1: Influence of Different Color Temperatures
Source: (Gao, 2024)

Research suggests that lighting conscious can significantly affect audience engagement and emotional response. According to Jain (2026), variations in stage lighting can affect viewers' emotional response during theatrical performances. A similar argument was also presented by Graham (2018), that specific lighting effects can act as a scenographic element that can communicate mood and meaning emotionally rather than simply illuminating performers and scenery. By carefully implementing specific colors and temperatures according to the mood of the performance, designers can shape how audiences interpret dramatic moments. Furthermore, Evans (2023) also highlighted the importance of audience emotions in theatrical experiences. Whereas Pelowski et al. (2019) found that lighting conditions have the capability to affect emotional appreciation of artistic works. These studies narrow down the fact that temperature has a crucial role in conveying the mood and emotional variations of the performance to the perceiving audience.

2.5 Theatrical Lighting and Audience Perception

As discussed earlier, theatrical lighting indeed plays a crucial role in shaping audience perception, besides enhancing the color accuracy. As Sokari (2022) highlighted, effective lighting design guides viewers toward important actions on stage and influences how scenes are interpreted. Hossain & Yadav (2025) also support that and address how lighting contributes significantly to the visual and dramatic quality of theatrical productions. Especially in musical theatre, where costumes, scenery, music, and performance contribute to the richness of visual

presentation, better lighting further creates an immersive experience for the audience. Supporting this, researchers like Li (2025) and Gómez-Sirvent et al. (2024) argue that environmental factors such as lighting can create psychological spaces that support the dramatic meaning of a performance, influencing audience perception and experience.

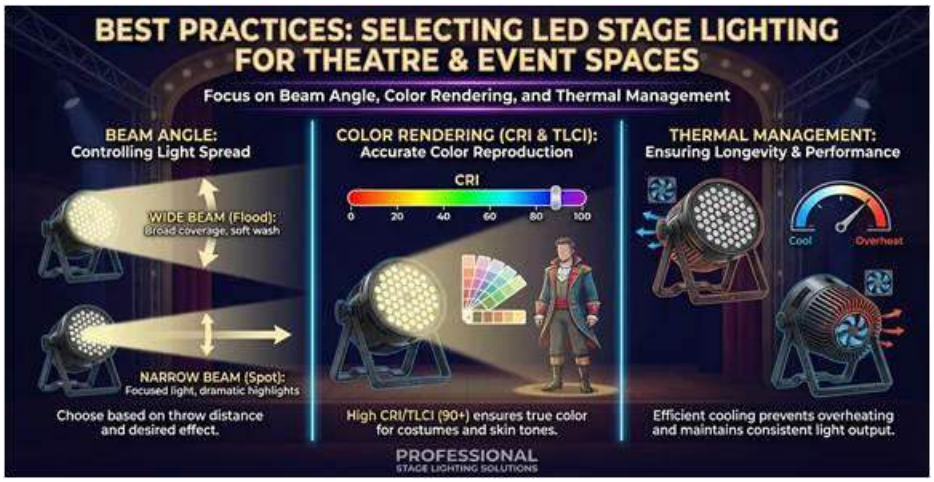


Figure 2.5.1: LED Placement Design
Source: (Zhang, 2025)

Researchers present the importance of the technical and design elements of light in musical theater to enhance overall audience engagement, as Figure 2.5.1 suggests. For example, Hashim et al. (2023) emphasize that technical design elements, including lighting effects, help strengthen the overall production and audience engagement. Zheng et al. (2025) further suggest that stage design elements influence emotional experiences and audience responses. The collective views from the researchers have indicated that theatrical lighting is not a mere vision medium but also a technical and artistic tool that shapes audience attention, emotional engagement, and overall perception.

2.6 Literature Gap

While all these studies have collectively examined LED lighting, color rendering, color temperature, and audience emotions, their discussion remains detached as they did not involve the mentioned factors at a time. Also, there is a lack of studies exploring how variations in LED spectrum characteristics simultaneously influence color perception, emotional engagement, and overall audience experience, particularly in musical theatre.

Chapter 3: Methodology

3.1 Research Design

This study employs a comparative experimental design to investigate how LED lighting characteristics influence audiences' perception and emotional response during theatrical performance. This research focuses specifically on two key properties of LED, which are color rendering and color temperature. Furthermore, this study compared the audiences' reactions to

the same theatrical environment with different LED lighting conditions. It explores whether changes in lighting conditions can affect both the visual appearance and emotional experience of the theatrical performance.

The comparative analysis approach is found to be more suitable for this study. It allows the influence of various lighting conditions to be deeply examined while keeping other aspects of the performance constant. As audiences' perception is inherently subjective, collecting responses directly from viewers would provide valuable insights into individual viewers' experiences.

3.2 Lighting Conditions Used for the Study

Two contrasting lighting conditions had been selected for the experiment. The sole purpose of using two conditions is to examine the differences in color rendering and color temperature and how they can influence audiences' overall perception.

In Lighting Condition A, a warm-white LED had been used, with an estimated color temperature of around 3000K and a CRI above 90. High CRI lighting is generally considered capable of reproducing more accurate colors of the subject. This makes the skin tone of the performers, the colors of the costumes, and the stage elements appear more natural (Led Lucky, 2022). And the warm color temperature is added as it is associated with feelings of comfort, intimacy, and emotional warmth.

In Lighting Condition B, a cool-white LED was utilized with an approximate color temperature between the range of 5000K and 6000K. In this setup, the cooler LED lighting is considered brighter, more energetic, or more clinical. As well as the cooler color temperature is used to create a contrast with the previous lighting condition. This helps to align and compare the audiences' visual reactions to the same theatrical scene.

Furthermore, to ensure a fair comparison, the brightness of both lighting scenarios was adjusted to similar levels as closely as possible. Other elements of the performance, such as actors, costumes, props, and stage arrangement, were kept unchanged during the experiment.

3.3 Participants and Procedure

To conduct the study successfully, a total of fifteen participants were recruited, including classmates, teachers, and other staff members. Participation of the candidates was voluntary, and all individuals agreed to take part in the survey. This sample was chosen because it represented a readily available audience that was suitable for conducting the study.

The participants were required to watch the same three-minute theatrical performance twice. The first scene was represented using Lighting Condition A. After the performance was finished, participants were provided with a questionnaire aligned with the topic of the study. This questionnaire assessed the participants regarding the scene, such as visual appearance, emotional engagement, and the quality of color. The procedure was repeated after the completion of the second performance with Lighting Condition B.

3.4 Questionnaire and Data Collection

Questionnaire responses were recorded using a five-point Likert scale ranging from 1 to 5, referring from strongly disagree to strongly agree, respectively. The Questionnaire, participants' data, and their averages are presented in the appendix section.

3.5 Data Analysis Method

The data accumulated from the questionnaire were analyzed using descriptive statistical methods. Responses were rated by implementing Likert Scale numerical values, ranging from 1 to 5. In the scale, 1 represented "Strongly Disagree" and 5 represented "Strongly Agree". Using the numerical value from the questionnaire, a mean score was derived for each statement. These participants were needed to rescore the questions after the completion of the theatrical performance in Lighting Condition B.

The results from the participants were organized into tables to facilitate the comparison between setups A and B. Additionally, ratings of the factors such as color perception, emotional engagement, atmosphere, and overall enjoyment of the performance were particularly scrutinized. The overall preference of the audience was also summarized using frequencies and percentages. As well as additional comments from the participants were reviewed to analyze common themes and observations regarding the effectiveness of each lighting condition.

3.6 Ethical Considerations

Several ethical conditions were taken into account during this study. All participants involved in this study were members of the school community, including students, teachers, and staff. Participation of each candidate was voluntary, and they were given detailed information about the intention and purpose of the experimentation before completing the questionnaire. Participants were also free to withdraw from the research at any time without any consequences. The experimentation was conducted for the sole purpose of the study. Therefore, information that was irrelevant to the study was not included, such as names, age, gender, occupation, and other personal information. This also helped to secure and protect the privacy of each participant. Also, the study was designed after ensuring that participants experienced no physical or psychological harm, and all collected data were analyzed confidentially.

Chapter 4: Result analysis

4.1 Color Perception Results

The first objective of the conducted experiment was to investigate how different LED lighting conditions can affect an audience's perception of the theatrical performance. For the experiment, the participants were provided with a 5-point Likert scale to score the questionnaire after their own observation of the theatrical performance. Later, a mean score was derived from the accumulation of results received from the participants.

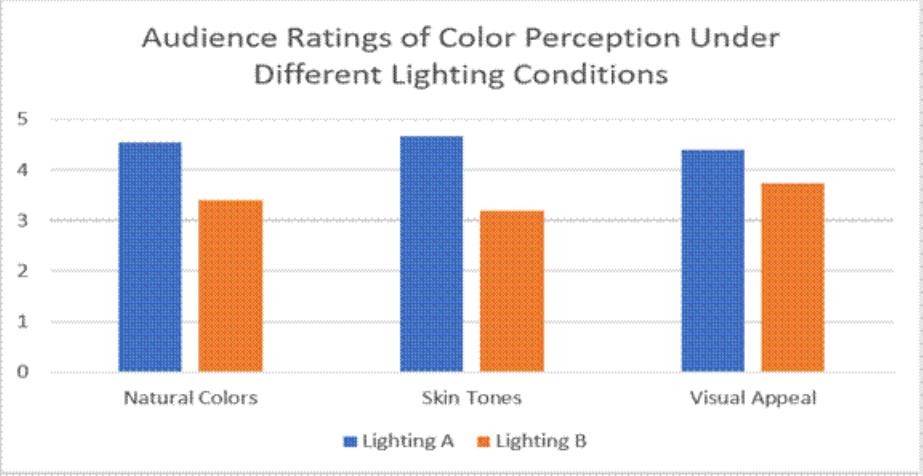


Figure 4.1.1: Audience Ratings of Color Perception Under Different Lighting Conditions

The questionnaire results revealed noticeable differences in audiences' perceptions of color quality and preferences between the two lighting conditions. For the questionnaire dealing with the accuracy of the reproduced natural colors of costumes and stage elements, Lighting Condition A scored higher with a mean of 4.53 than B, which received only 3.4. Moving on, the next questionnaire representing the accuracy of the performer's skin tone, Lighting setup A scored 4.67 compared to the score of 3.2 for Lighting B. A similar pattern was also observed for the questionnaire regarding the overall appealing factor for the performance. Lighting setup A was able to score 4.4, whereas Lighting B was only able to accumulate a score of 3.73.
As per the data shown in Figure 4.1.1, Lighting Condition A provided superior visual quality in comparison to setup B. By combining warm color temperature and a high CRI value enhances the realism and overall attractiveness of the theatrical performance and its environment.

4.2 Emotional Response Results

After deriving the first set of results, the study progressed to its second objective. It was to examine how LED lighting conditions can influence audiences' emotional engagement and benefit the overall experience of the theatrical performance.

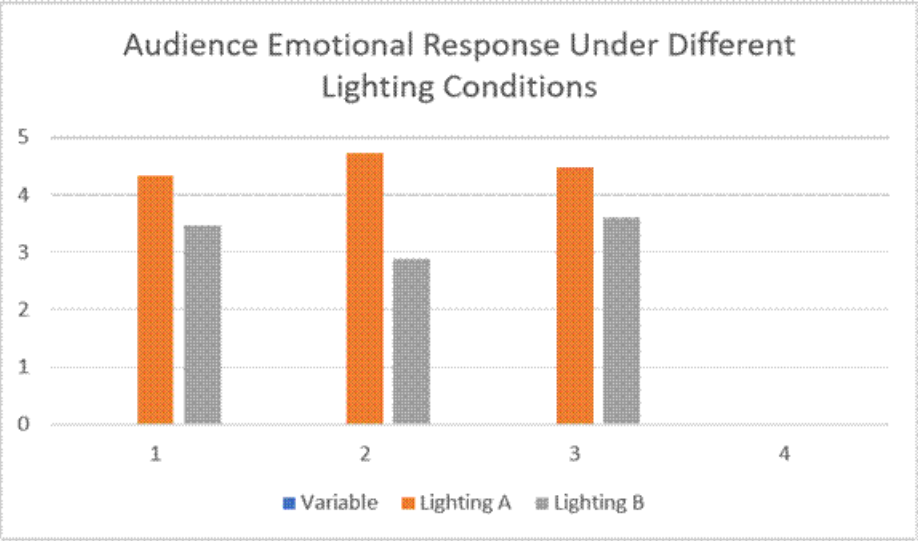


Figure 4.2.1: Audience Emotional Response Under Different Lighting Conditions

The statement "The lighting helped me connect emotionally with the performance" received an accumulated score of 4.33 under Lighting A, whereas Lighting B scored 3.47. This suggests that warmer lighting conditions can strengthen audiences' emotional engagement and increase their immersion into the flow of the theatrical performance. Moreover, the largest difference in score was found in the results of the statement "The atmosphere felt warm and inviting". In this statement, Lighting A was able to score 4.7, whereas Lighting B scored only 2.83. This difference in points between Lighting Condition A and B proves that color temperature has an impact on audiences' perception of mood and atmosphere. The majority of the participants chose the warm light setup over the one with the cooler temperature, because it associated easily with the feelings of comfort and emotional warmth. A similar pattern occurred in the statement "I enjoyed watching the scene under this lighting", where the majority of the score went to Lighting A. It was able to accumulate an average score of 4.47, which exceeded the Lighting B score of

3.6. These findings suggest that improved visual quality along with a warmer atmosphere contributes positively to the audience's enjoyment.

In general, the importance of emotional response results was to find which lighting setup has the most impact in bringing out the overall quality of the performance. Lighting A was able to achieve higher scores than Lighting B and proved to be more effective in creating a comfortable, engaging, and emotionally meaningful theatrical experience.

4.3 Comparative Analysis

In addition, participants were asked to input their overall preferred lighting setup after viewing both performances.

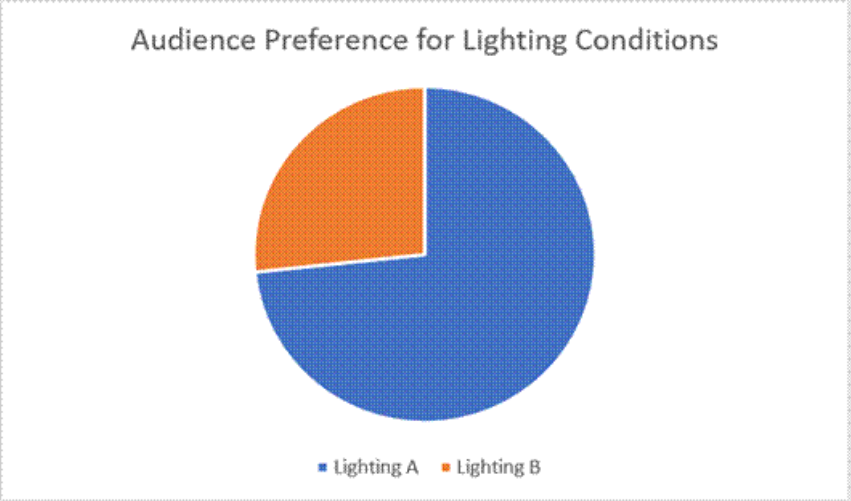


Figure 4.3.1: Audience Preference for Lighting Conditions

Figure 4.3.1 highlights the clear distinction in the audience's preferences regarding the Lighting setups. Lighting Condition A was able to achieve 11 votes, whereas Lighting Condition B was only able to secure 4. These results verify the findings showcased in Figures 4.1.1 and 4.2.1. The consistency in superiority of Lighting Condition A across both visual and emotional categories suggests that audiences respond highly to lighting setups similar to Lighting Condition A. This setup allows participants to perceive the overall performance, and also other details of it, such as costumes, skin tones of the performers, and stage elements, better.

In a broader perspective, the results indicate that LED spectrum characteristics can contribute more toward audiences' strong emotional engagement and increase the enjoyment of the performance. These findings support the proposed hypothesis that a high CRI value and warmer color temperature can positively impact the visual quality and emotional response of the audience during a theatrical performance.

Chapter 5: Discussion

The findings of the study showcase that LED lighting characteristics, such as color rendering and color temperature, can influence audience perception during theatrical performance. Analysis of the results indicates the positive effectiveness of Lighting Condition A in maintaining the naturalness of colors, skin tones, emotional engagement, and overall enjoyment. These outcomes support the proposed hypothesis that warm color temperature and high CRI value can increase the experience of the theatrical performance.

The results also align with the research discussed in the literature review. Studies by Lee & Yoon (2021) and Siniscalco (2022) point out the importance of color rendering to improve the overall visual quality. Whereas Jain (2026) and Graham (2018) demonstrated that lighting conditions can affect audiences' emotional response during theatrical performance. Similarly, Sokari (2022) also highlighted how lighting influences audiences' attention and perception during performance.

Through a scientific viewpoint, these findings can be explained through the interaction of light spectra and human perception. LEDs with a high CRI value can reproduce the reflected wavelengths more accurately, allowing costumes, scenery, and skin tones to appear more natural. And at the same time, warm color temperatures are mostly associated with feelings of comfort and intimacy, which may explain the stronger emotional engagement reported by participants. Therefore, school theatre productions are highly recommended to utilize LEDs with high CRI values above 90 when possible. Furthermore, warm lighting can also be highly effective for emotionally driven scenes. However, careful selection of color temperature can also help to boost the creativity and express the emotion of the performance while maintaining a strong audience immersion and storytelling.

Chapter 6: Limitations

Across this study, there are several limitations that need to be considered during the interpretation of findings. First, the sample size was relatively small, consisting only of fifteen members selected from the school community. This may limit the applicability of the results. Second, the study examined only one theatrical performance with two lighting conditions, whereas in real productions, a large variety of lighting effects and colors are used during the performance. Third, audience responses were subjectively measured using a simple questionnaire. Therefore, future research with larger audiences, multiple performances, and various lighting configurations can help to derive more conclusive results.

Chapter 7: Conclusion

In conclusion, this study investigated how LED lighting characteristics, such as color rendering and color temperature, can influence audience perception and emotional response during a theatrical performance. These findings show that warm temperature and a high CRI value lighting setup can help the audience to perceive the naturalness of the performers, colors, and other stage elements. The findings of the research also suggest that optimizing the spectrum characteristics of LED lighting setups can directly influence the visual quality and impact on the performance of audiences. Therefore, effective lighting is not only a technique just to illuminate the character for the sake of visibility but rather to draw the raw emotions of the performance.

Works Cited

ABS. (2022, August 28). Important Stages of Emotional Healing from a Holistic Perspective. Retrieved from A Beautiful Soul Holistic Counseling website: https://beautifulsoulcounseling.com/important-stages-of-emotional-healing/

Alrajhi, A. A., Alsahli, A. S., Alhelal, I. M., Rihan, H. Z., Fuller, M. P., Alsadon, A. A., & Ibrahim, A. A. (2023). The effect of LED light spectra on the growth, yield and nutritional value of red and green lettuce (Lactuca sativa). Plants, 12(3), 463. https://www.mdpi.com/2223-7747/12/3/463

Asuquo, N. B. (2025). Towards An Era Of Eco-Friendly Theatre Lighting Equipment: An Imperative For A Transition From Incandescence To Light-Emitting-Diodes. Journal of Emerging Trends in Educational Research and Policy Studies, 16(2), 38-45. https://journals.co.za/doi/abs/10.10520/ejc-sl_jeteraps_v16_n2_a2

Barrett, E. M., & Jeffery, G. (2026). LED lighting (350-650nm) undermines human visual performance unless supplemented by wider spectra (400-1500nm+) like daylight. Scientific Reports, 16(1), 3061. https://www.nature.com/articles/s41598-026-35389-6

Basa, M. (2023). Lighting design schemes and colors in dance performances: The magical illusion. ShodhKosh: Journal of Visual and Performing Arts, 4(1), 356-362. https://www.researchgate.net/profile/Murali-Basa/publication/370492412_LIGHTING_DESIGN_SCHEMES_AND_COLOURS_IN_DANCE_PERFORMANCES_THE_MAGICAL_ILLUSION/links/6452d44b4af7887352572812/LIGHTING-DESIGN-SCHEMES-AND-COLOURS-IN-DANCE-PERFORMANCES-THE-MAGICAL-ILLUSION.pdf

Bell, A. (2013). The Effect of Colour on Audience Response in Theatre Scenic Design. https://www.researchgate.net/profile/Allison-Bell-7/publication/378403177_The_Effect_of_Colour_on_Audience_Response_in_Theatre_Scenic_Design/links/65d883d5adf2362b6352d887/The-Effect-of-Colour-on-Audience-Response-in-Theatre-Scenic-Design.pdf

Bishop, D., & Chase, J. G. (2023). A luminance-based lighting design method: A framework for lighting design and review of luminance measures. Sustainability, 15(5), 4369. https://www.mdpi.com/2071-1050/15/5/4369

Cao, M., Luo, M. R., & Liu, G. (2020, November). Preferred skin tones reproduction of three ethnic groups under different ambient lighting conditions. In Color and Imaging Conference (Vol. 28, pp. 94-99). Society for Imaging Science and Technology. https://library.imaging.org/cic/articles/28/1/art00014

Chan, A., & Cheung, N. (2022). „Reimagining Light, Technology and Spectatorship in Contemporary Theatre". Critical Stages, 25. https://www.critical-stages.org/25/reimagining-light-technology-and-spectatorship-in-contemporary-theatre/

Evans, R. (2023). Catching feelings: Measurement of theatre audience emotional response through performance (Doctoral dissertation, University of Central Lancashire). https://knowledge.lancashire.ac.uk/id/eprint/46180/

Gao, T. (2024, August 31). What is RGB Color? Retrieved from Epackprinting.com website: https://www.epackprinting.com/support/what-is-rgb-color/

Gómez-Sirvent, J. L., Fernández-Sotos, A., Fernández-Caballero, A., & Fernández-Sotos, D. (2024). Assessment of music performance anxiety in a virtual auditorium through the study of ambient lighting and audience distance. Virtual Reality, 28(2), 71. https://link.springer.com/article/10.1007/s10055-024-00947-8

Graham, K. J. (2018). Scenographic light: towards an understanding of expressive light in performance (Doctoral dissertation, University of Leeds). https://etheses.whiterose.ac.uk/id/eprint/20414/

Hanlon, K. L., Wei, G., Correa-Selm, L., & Grichnik, J. M. (2022). Dermoscopy and skin imaging light sources: a comparison and review of spectral power distribution and color consistency. Journal of biomedical optics, 27(8), 080902-080902. https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-27/issue-8/080902/Dermoscopy-and-skin-imaging-light-sources--a-comparison-and/10.1117/1.JBO.27.8.080902.short

Hashim, H., Zulkepeli, M. K., Abd Rahman, K. A. A., & Siu, K. W. M. (2023). Technical Design and Production of Special Effects in Musical Theatre Performances. Journal of ASIAN Behavioural Studies, 8(24), 67-78. http://jabs.e-iph.co.uk/index.php/jABs/article/view/424

Hossain, M. R., & Yadav, C. (2025). The Impacts of Theatre Lighting in Play Productions. Vol. XXI, 202. https://ssei.org.in/TI/Vol-XXI.pdf#page=223

Hurlbert, A., & Yu, C. (2025). Seeing the light: Perception and discrimination of illumination color. Annual Review of Vision Science, 11. https://www.annualreviews.org/content/journals/10.1146/annurev-vision-121423-013755

Jain, D. (2026). Psychological effects of dynamic stage lighting on audience engagement in theatre. The Criterion: An International Journal in English, 17(2), 1007–1030. https://doi.org/10.66376/criterion.v17.n2.64

Kashaka, N. D. (2025). Cinematic techniques: Shaping viewer perception and emotion. Research Invention Journal of Current Issues in Arts and Management, 4(2), 28-33. https://rijournals.com/wp-content/uploads/2025/02/RIJCIAM-42-2025-P7.pdf

Led Lucky. (2022, November 11). The Comprehensive Guide to Meeting Room Lighting. Retrieved from Ledlucky website: https://ledlucky.net/the-comprehensive-guide-to-meeting-room-lighting/

Lee, H., & Lee, E. (2025). An Experimental Investigation into the Influence of Colored Lighting on Perceived Spatial Impressions. Buildings, 15(19), 3511. https://www.mdpi.com/2075-5309/15/19/3511

Lee, S., & Yoon, H. C. (2021). LED lighting system for better color rendition space: the effect of color rendering index. Journal of Asian Architecture and Building Engineering, 20(5), 556-565. https://www.tandfonline.com/doi/abs/10.1080/13467581.2020.1799801

Li, A. (2025). An Exploration of Psychological Space-Time in Theatre and Its Manifestation through Stage Lighting. Asian Arts and Society Journal, 1(2), 17-38. https://so13.tci-thaijo.org/index.php/AASJ/article/view/3138

Liang, S. (2025, April 19). What is the most important function for stage lighting? - Vorlane. Retrieved from Vorlane website: https://vorlane.com/most-important-function-stage-lighting/

Lifocolor. (2025, June 2). How colors are created and why we see them differently. Retrieved from Lifocolor.de website: https://www.lifocolor.de/en/news-events/why-people-perceive-colours-differently

Marlow, P. J., Gegenfurtner, K. R., & Anderson, B. L. (2022). The role of color in the perception of three-dimensional shape. Current Biology, 32(6), 1387-1394. https://www.cell.com/current-biology/fulltext/S0960-9822(22)00037-9

McDonald, J., Canazza, S., Chmiel, A., De Poli, G., Houbert, E., Murari, M., ... & Zhang, J. D. (2022). Illuminating music: Impact of color hue for background lighting on emotional arousal in piano performance videos. Frontiers in Psychology, 13, 828699. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.828699/full

Miller, N. J., Kaye, S., Coleman, P., Wilkerson, A. M., Perrin, T. E., & Sullivan, G. P. (2014). LED Lighting in a Performing Arts Building at the University of Florida (No. PNNL-23514). Pacific Northwest National Laboratory (PNNL), Richland, WA (US). https://www.osti.gov/biblio/1165331

NASA. (2023). Visible Light. Retrieved from science.nasa.gov website: https://science.nasa.gov/ems/09_visiblelight/

NASA. (2024). NASA PACE - Section I: Light Properties. Retrieved from Oceansciences.org website: https://pace.oceansciences.org/prop_of_light.cgi

Oh, S. T., & Lim, J. H. (2023). CRI-Based Smart Lighting System That Provides Characteristics of Natural Light. Information, 14(12), 628. https://www.mdpi.com/2078-2489/14/12/628

Oliveira, A. P., Ferreira, G., & Martinez-Perez, C. (2024). Effects of Stage Lighting on Visual Comfort at Summer Festivals: A Study in Portugal. Healthcare (Basel, Switzerland), 12(23), 2441. https://doi.org/10.3390/healthcare12232441

Pattison, P. M., Hansen, M., & Tsao, J. Y. (2018). LED lighting efficacy: status and directions. Comptes Rendus Physique, 19(3), 134-145. https://www.sciencedirect.com/science/article/pii/S1631070517300932

Pelowski, M., Graser, A., Specker, E., Forster, M., von Hinüber, J., & Leder, H. (2019). Does gallery lighting really have an impact on appreciation of art? An ecologically valid study of lighting changes and the assessment and emotional experience with representational and abstract paintings. Frontiers in Psychology, 10, 2148. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02148/full

Prangnell, L. (2016). Visible light-based human visual system conceptual model. arXiv preprint arXiv:1609.04830. https://arxiv.org/abs/1609.04830

Rossi, S., Kara-José, N., Rocha, E. M., & Kara-Júnior, N. (2024). Influence of lighting on visual performance. Arquivos brasileiros de oftalmologia, 87(3), e20230257. https://doi.org/10.5935/0004-2749.2023-0257

Sarinastiti, W., Murdaningtyas, C. D., & Sopannadi. (2024, October). The interplay of lighting and graphics in musical performances concert. In AIP Conference Proceedings (Vol. 3222, No. 1, p. 040009). AIP Publishing LLC. https://pubs.aip.org/aip/acp/article-abstract/3222/1/040009/3316266

Shahidi, R., Golmohammadi, R., Babamiri, M., Faradmal, J., & Aliabadi, M. (2021). Effect of warm/cool white lights on visual perception and mood in warm/cool color environments. EXCLI journal, 20, 1379–1393. https://doi.org/10.17179/excli2021-3974

Shuaib, S., & Mohammed, R. J. (2026). Technical Analysis of Integrated Led and Automated Lighting Systems in Nigerian Theatrical Productions: A Comparative Case Study of Saro The Musical and Death and The King's Horseman. Nigerian Journal of Technological Development, 23(1), 388-394. https://journal.njtd.com.ng/index.php/njtd/article/view/4686

Siniscalco, A. (2022). A design approach to lighting and color rendering in indoor sets. Journal of the International Colour Association, 29. https://www.aic-publishing.org/ojs/index.php/JAIC/article/view/257

Smith, D. A., & Coleman, B. (2021). Lights, physics, action: The science of theater lighting design. The Physics Teacher, 59(8), 602-604. https://pubs.aip.org/aapt/pte/article/59/8/602/278851

Sokari, I. F. (2022). The Effect Of Lighting Design On Audience Attention.https://www.researchgate.net/profile/Ibifuro-Sokari/publication/389659419_THE_EFFECT_OF_LIGHTING_DESIGN_ON_AUDIENCE_ATTENTION/links/67cbb35be62c604a0dd637ac/THE-EFFECT-OF-LIGHTING-DESIGN-ON-AUDIENCE-ATTENTION.pdf

Sokari, I. F. (2025). DESIGN EFFECT IN THE THEATRE: ENHANCING THE AUDIENCE EXPERIENCE. https://www.researchgate.net/profile/Ibifuro-Sokari/publication/389737423_DESIGN_EFFECT_IN_THE_THEATRE_ENHANCING_THE_AUDIENCE_EXPERIENCE/links/67d02f2c7c5b5569dcb9776b/DESIGN-EFFECT-IN-THE-THEATRE-ENHANCING-THE-AUDIENCE-EXPERIENCE.pdf

Take Three Lighting. (2026). Understanding CRI & Why It's Important with LED Lights. Retrieved from www.takethreelighting.com website: https://www.takethreelighting.com/understanding-cri.html

Upward Lighting. (2023). 3000K vs. 4000K vs. 6000K: Which Lighting Is Suitable For Home? Retrieved from Upward Lighting website: https://upwardlighting.com/3000k-vs-4000k-vs-6000k/

Wu, L., Li, Y., Chen, H., Wang, J., Wang, X., & Xue, Y. (2026). The spectral power distribution prediction of LED light source based on Gaussian mathematical model and improved

residual network. Scientific Reports. https://www.nature.com/articles/s41598-026-39015-3

Xie, X., Cai, J., Fang, H., Tang, X., & Yamanaka, T. (2022). Effects of colored lights on an individual's affective impressions in the observation process. Frontiers in Psychology, 13, 938636. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.938636/full?utm_source=chatgpt.com

Zhang, C., Halabi, K. N. M., Azahari, M. H., & Wu, Y. (2023). Analysis of the design and use of stage lighting and artistic expressions. Art and Performance Letters, 4(9), 33. https://www.researchgate.net/publication/374373142_Analysis_of_the_design_and_use_of_stage_lighting_and_artistic_expressions

Zhang, L. (2025, December 30). How To Choose LED Stage Lighting For Theatres, Churches, Clubs, And Events. Retrieved from SignliteLED website: https://www.signliteled.com/choose-led-stage-lighting-theatres-churches-clubs-events

Zhao, E. Z., Dong, L. L., Chen, Y., Lou, Q., & Xu, W. H. (2020). The impact of LED color rendering on the dark adaptation of human eyes at tunnel entrances. International journal of environmental research and public health, 17(5), 1566. https://www.mdpi.com/1660-4601/17/5/1566

Zheng, J., Ye, T., Zhan, C., Li, Y., & Bi, Y. (2025)The Impact of Stage Design on Audience Emotional Experience: A Perspective from Drama and Film Studies. Available at SSRN 5154590. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5154590

Zhou, B., Zhou, Y., Ren, Q., Peng, L., Guan, Y., & He, H. (2025). Research on the Impact of Lighting Illuminance and Color on Creative Performance and Mood. Buildings, 15(10), 1738. https://www.mdpi.com/2075-5309/15/10/1738

Self-Perception, Well-being, and Collaborative Decision-Making in Middle School Students: A Cross-Sectional Study by Grade and Gender By Stella Kakkanatt

Abstract

Although collaboration has been widely studied, little is known about what predicts a student's decision to collaborate in the first place. Because social connection is essential to both individual well-being and the functioning of social groups, perhaps well-being and collaboration are linked. We designed this study to explore the relationship between well-being traits and the real-time choice to collaborate in middle school students. We surveyed 177 middle school students on well-being traits and preference for working with others, then presented them with a real-time anonymous choice: share puzzle answers with another student, or work alone. We hypothesized that students with more positive well-being and social preferences would be more likely to share. Our hypothesis was partially supported: well-being traits did not predict the sharing decision, with the exception of health in seventh and eighth graders. Instead, stated preference for collaboration was the strongest predictor of sharing, especially in older girls ($r=0.51$, $p=0.007$). To better understand this shift, we examined how well-being traits related to each other. Both genders showed similar increases in the differentiation of well-being traits with age, suggesting that adolescents develop more complex self-perception as they mature. However, the pattern of differentiation differed between genders. Girls' well-being traits stayed almost completely positively correlated throughout, while boys showed an increasing number of negative correlations between traits. Together, these findings suggest that gender shapes both the development of preference-behavior connection and the structure of well-being in adolescence. These findings can help inform well-being program design and collaboration in education.

Introduction

Adolescent well-being is multidimensional, and includes health, self-perception, social connection, and resilience (Ross et al.). Research suggests that programs targeting one well-being trait tend to improve others, suggesting that these measures reinforce each other rather than moving independently (Ross et al.; McLoughlin and Kristjansson; Verma). Longitudinal research shows that social connection is essential to well

being. The longest-running study of human happiness, the Harvard Study of Adult Development, found that the quality of close relationships predicts health outcomes, happiness, and even longevity (4). Even brief social interactions such as greetings and casual conversation have been shown to predict life satisfaction (Sandstrom and Dunn; Ascigil et al.). If connection is fundamental to well-being, then the decision to connect or collaborate may be an important part of healthy adolescent development.

However, the research on adolescent collaboration has focused mostly on the quality and outcomes of collaboration after students have been put in groups. Gauvain and Rogoff found that children assigned to collaborate performed planning tasks more effectively than those who

worked alone. Swenson and Strough found that the quality of the relationships between assigned partners predicted collaborative success, and that girls slightly outperformed boys, but again, participants were not given the choice to collaborate. No prior study has investigated whether well-being traits predict the real-time decision to collaborate, nor how that relationship might develop across middle school grade levels.

This gap matters because middle school is a time of rapid development in self perception and social behavior (Ross et al.; Verma). Self-perception becomes more differentiated between early and late adolescence. Gender differences in collaborative behavior have been described (Swenson and Strough) but the factors influencing these differences, especially well-being and preference, have not been fully studied. Understanding whether and how well-being predicts the choice to collaborate, and how that changes across grades and genders, could help educators create more effective wellness programs and collaborative classroom environments.

We designed this study to fill this gap. Unlike previous studies that assigned students to collaborative groups, we measured collaboration as a behavioral choice. We hypothesized that students reporting more positive self-perception, social preferences, and well-being would be more likely to share answers during a challenging task. Our hypothesis was only partially supported. Preference, not well-being, was the strongest predictor of sharing behavior, and the relationship between stated preference and behavior strengthened dramatically in older girls but not in boys. Finally, we found that both genders showed increasing differentiation of well-being traits with age, but diverged in the direction of inter-trait relationships, with boys showing a growing number of negatively correlated traits.

Results

Sharing rates differed by grade and gender

Approximately half of all participants chose to share their puzzle solutions (Figures 1,2). Seventh and eighth graders shared more than sixth graders, and this difference was almost entirely due to the girls. Among the sixth grade girls, 36% chose to share, compared to 57% of seventh and eighth grade girls, a 21 percentage point increase. In contrast, sixth grade boys and seventh/eighth grade boys shared at statistically identical rates (48% vs. 50%), reflecting no significant change across grade levels.

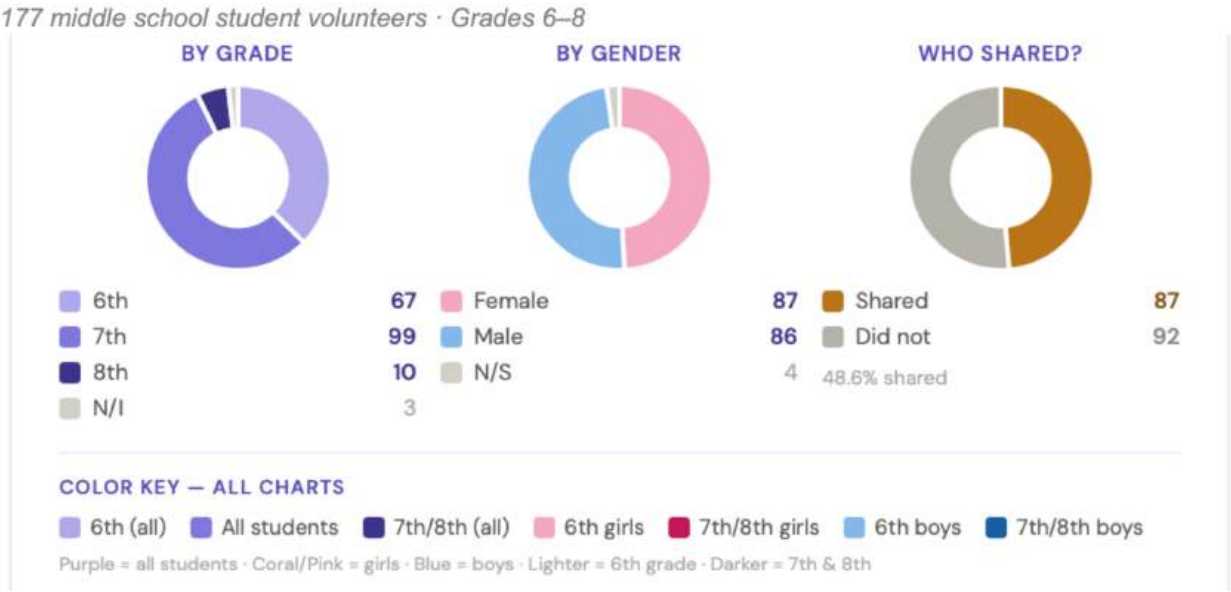


Figure 1. Study participants and overall sharing rates. Donut charts showing the composition of study participants by grade and gender (n = 177) and the proportion of all participants who chose to share puzzle answers versus work alone.

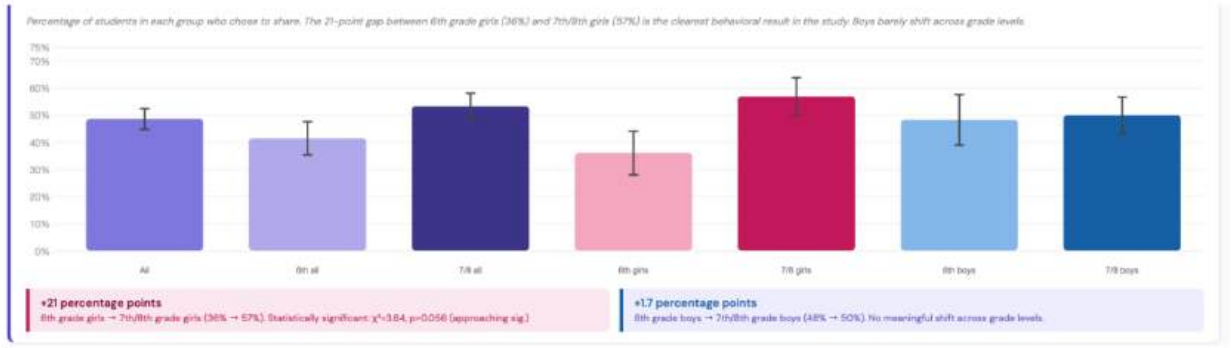


Figure 2. Sharing rates by grade and gender. Bar graph showing the percentage of students who chose to share puzzle answers, broken down by grade level (sixth grade vs. seventh/eighth grade) and gender (girls vs. boys). Sixth grade girls shared at a lower rate than seventh/eighth grade girls (36% vs. 57%), while boys showed no meaningful change across grade levels (48% vs. 50%).

Work preference, not well-being, predicted the sharing decision

To measure the correlation between survey responses and the sharing decision, we calculated Pearson correlations between each survey measure and the decision behavior for the full sample, as well as for the subgroups by grade and gender. For all groups, work preference

was the strongest predictor of sharing (r=0.13 to 0.51 across subgroups), with students preferring to work with others more likely to share (Figure 3-5). Health was the only well-being trait that showed a statistically significant positive correlation with sharing, and only in seventh and eighth graders overall (r=0.201, p=0.038). Academic confidence showed a slight negative association with sharing across all groups, though this did not reach statistical significance.

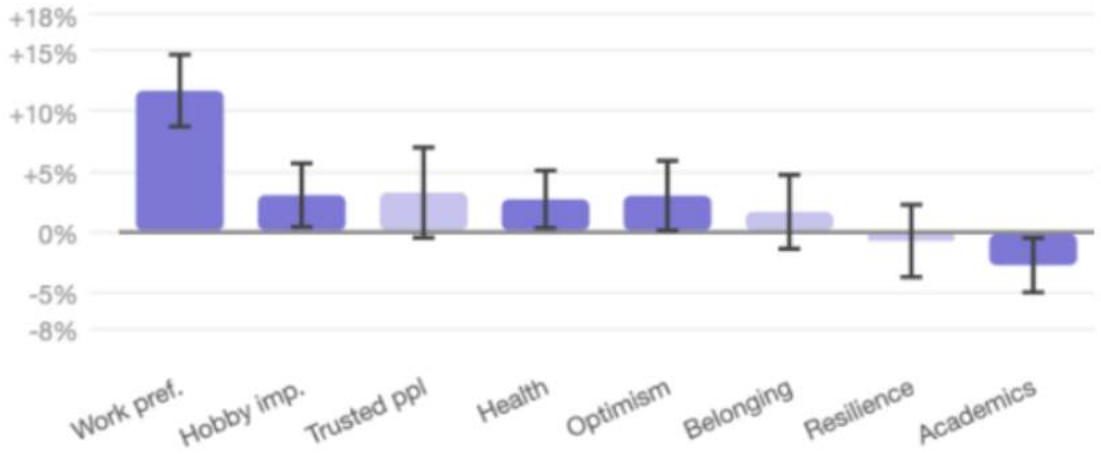


Figure 3. Pearson correlations between survey measures and sharing decision for all participants. Bar chart showing mean difference in Likert scores between sharers and non-sharers for each survey measure, with 95% confidence interval error bars. Work preference showed the strongest positive association with sharing behavior (r = 0.13 to 0.51 across subgroups).

In sixth grade, the error bars for all survey measures crossed zero, meaning no survey response, including work preference, reliably predicted the decision to share (Figure 4). By seventh and eighth grade, however, work preference strongly predicted collaborative behavior in girls (r=0.51, Figure 5). A Fisher's r-to-z test showed that the shift in correlation between work preference and sharing from sixth grade girls (r=-0.05) to seventh/eighth grade girls (r=0.51) was highly statistically significant (z=2.68, p=0.007, Figure 6). For boys, the same correlation was r≈0.30 at both grade levels, with no significant change (p=0.98).

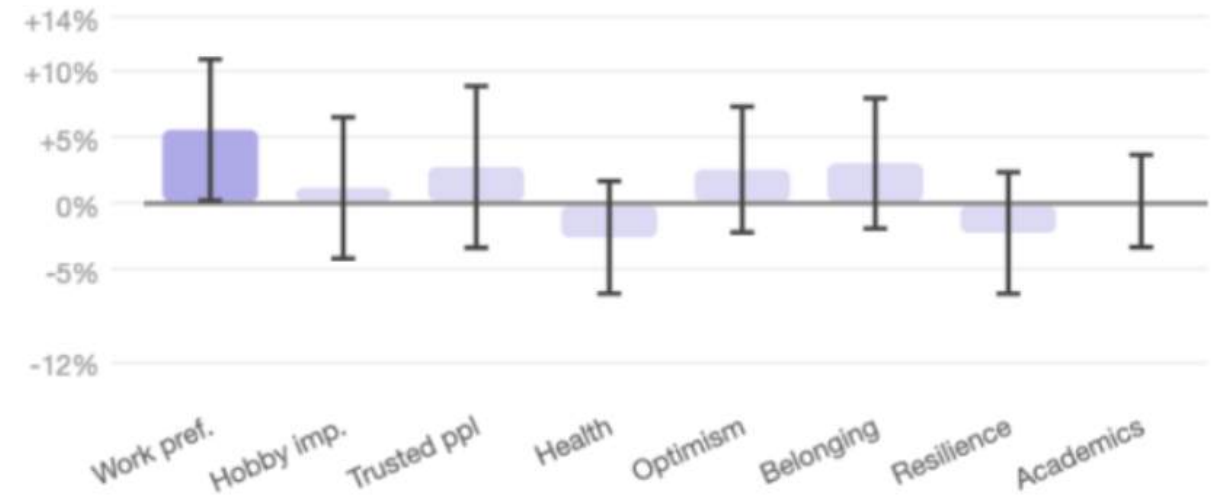


Figure 4. Survey measure correlations with sharing in sixth graders. Bar chart showing mean difference in Likert scores between sharers and non-sharers for sixth grade students, with 95% confidence interval error bars. Error bars for all measures cross zero, indicating no survey response reliably predicted the sharing decision in sixth grade.

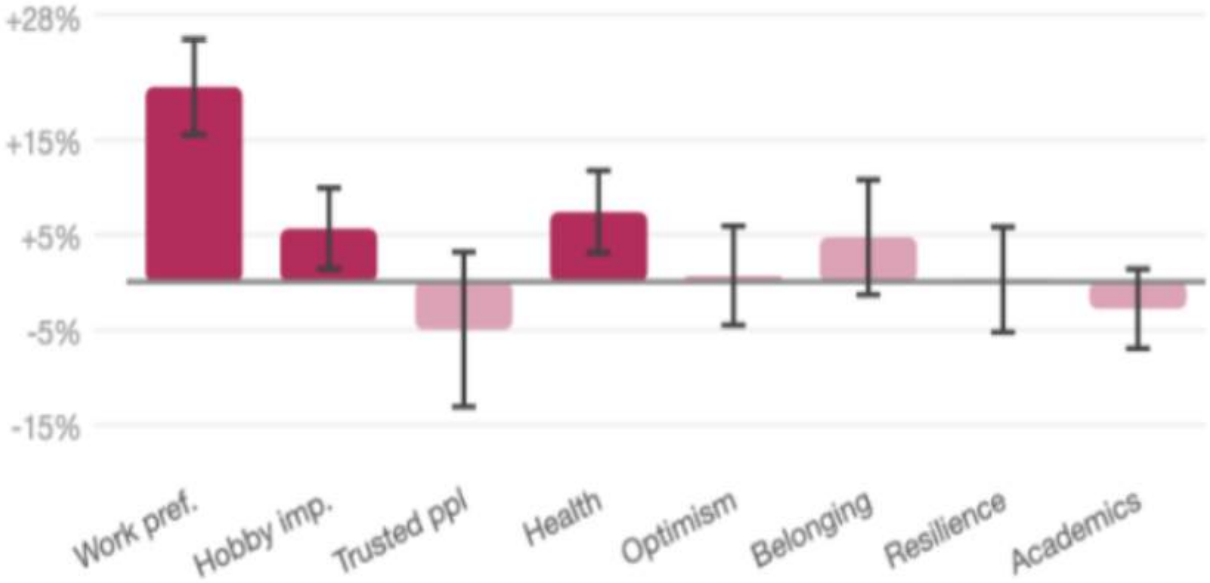


Figure 5. Survey measure correlations with sharing in seventh and eighth grade girls. Bar chart showing mean difference in Likert scores between sharers and non-sharers for seventh and eighth grade girls, with 95% confidence interval error bars. Work preference showed a strong positive association ($r = 0.51$), with the error bar clearly excluding zero.

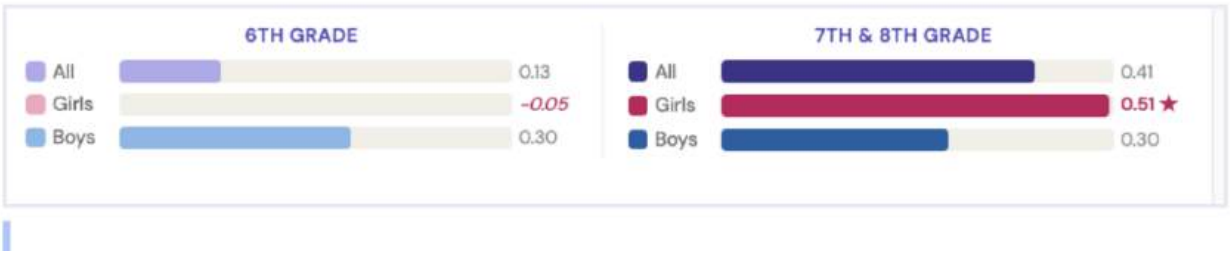


Figure 6. Grade-related change in correlation between work preference and sharing by gender. Bar chart comparing Pearson r values for the correlation between work preference and sharing decision across subgroups. The shift from sixth grade girls (r = -0.05) to seventh/eighth grade girls (r = 0.51) was statistically significant (z = 2.68, p = 0.007, Fisher's r-to-z test). Boys showed no significant change across grade levels (r ≈ 0.30 at both levels, p = 0.98).

Well-being traits de-clustered with age in both genders, but diverged in direction

We also used Pearson correlations to determine how well-being traits related to each other within subgroups. In sixth grade, students tended to rate themselves similarly across all well-being traits, whether high, low, or medium across the board, consistent with a general sense of overall well-being (mean inter-trait r [MIr]=0.235, Figure 7). By seventh and eighth grade, traits became more differentiated with the overall MIr dropping to 0.144 (Figures 7,8). Both genders de-clustered at similar rates: girls by 35.7% (MIr: 0.283 to 0.182) and boys by 33.7% (MIr:0.208 to 0.138).

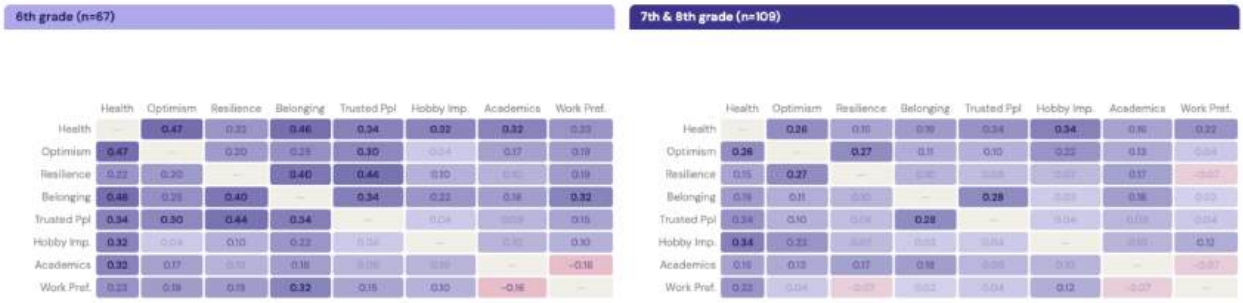


Figure 7. Inter-trait correlation heat map by grade level, both genders. Heat maps showing Pearson r values between all pairs of well-being traits for (7a) sixth graders (n = 65, MIr=0.235) and (7b) seventh and eighth graders (n=112, MIr = 0.144), reflecting de-clustering of well-being traits with age. MIr was calculated using absolute values.

However, the direction of inter-trait correlations differed between genders. MIr was calculated using absolute values to measure clustering strength regardless of direction. Girls trait pairs nearly all stayed positively correlated through middle school, with only 1 of 28 possible pairs negatively correlated at both grade levels. Boys began with 3 negative pairs in sixth grade

and increased to 7 of 28 pairs by seventh and eighth grade, meaning that a growing portion of boys' well-being traits moved in opposite directions as they matured (Figure 8). Health was the most consistently connected traits across all subgroups, showing the highest average correlation with other well-being traits including optimism, belonging, trusted people, and hobby importance.

Figure 8. Inter-trait correlation heat maps by grade and gender. Four heat maps showing Pearson r values for well-being trait pairs in sixth grade girls (MIr = 0.283, 1 of 28 pairs negative), seventh/eighth grade girls (MIr = 0.182, 1 of 28 pairs negative), sixth grade boys (MIr = 0.208, 3 of 28 pairs negative), and seventh/eighth grade boys (MIr = 0.138, 7 of 28 pairs negative). Negative pairs are shown in pink. Both genders de-clustered at similar rates (girls: 35.7%; boys: 33.7%), but boys showed a growing proportion of negatively correlated trait pairs with age.

Discussion

Our hypothesis was partially supported. With the exception of health in older students, well-being traits did not predict the decision to share answers. Instead, preference for collaboration was the strongest and most consistent predictor of the sharing behavior.

However, that relationship was not the same across all students. One of the most striking findings of this study was that three changes in girls occurred simultaneously. Between sixth and seventh/eighth grade, girls shared more, their preferences became stronger predictors of behavior, and their well-being traits became more differentiated. Together, these findings suggest a developmental change: as girls developed a more sophisticated sense of their own well-being, they may also have acted more consistently according to their preferences. These

patterns are consistent with developmental changes in self-perception, although the cross sectional design of this study cannot prove the underlying cause.

To better understand this developmental shift, we studied how students' well-being traits related to each other. This analysis described how students considered different aspects of their own well-being, rather than simply asking which traits predicted sharing.

Both genders showed similar rates of differentiation of well-being traits with age (35.7% for girls, 33.7% for boys), suggesting that increasing complexity of self-perception occurs in all adolescents. However the nature of differentiation diverged between genders. Girls' trait pairs stayed nearly all positively correlated throughout middle school (1 of 28 negative pairs at both grade levels), while boys showed an increase in negatively correlated pairs (3 to 7 of 28 between sixth and seventh/eighth grade). This pattern suggests that older boys viewed some well-being traits not only independently, but possibly in opposition against one another, rather than moving in the same direction. This finding has not been previously described and may represent a different pattern of development in self-perception that can be further studied.

These patterns in development go beyond prior research on adolescent well-being. Earlier studies describe well-being measures as reinforcing one another (1,3), a pattern we observed in sixth graders. However, this gender-specific pattern of de-clustering suggests that self-perception becomes more differentiated during middle school, with the correlation between traits developing differently for girls and boys. Prior studies by Swenson and Strough studied gender differences in collaboration between assigned partners (8), but this study adds a new angle by recording gender differences in the choice to collaborate, and the change in well-being trait relationships over different age groups.

Of all the well-being measures that we surveyed, self-perception of health was unique. In almost every subgroup, health was most linked to the other traits, particularly belonging, optimism, trusted people, and hobby importance. This fits with the findings of the Harvard Study of Adult Development, which found that social connection and health predict each other over time (Waldinger and Schulz), and with research showing that well-being dimensions reinforce one another (1). In addition, health was the only well-being trait to show a statistically significant association with the decision to share, but only in the seventh and eighth grade. This suggests that as students mature, how they perceive their physical health may become more connected to their behavior.

There were several limitations. This study sampled a single school, limiting how we can generalize the results. The observational design establishes correlation, but not causation. Surveys are prone to self-report bias, and the anonymous choice design, meant to decrease social anxiety, cannot fully erase social influence in the classroom. Some subgroups, especially eight graders (n=10), were too small for statistical analysis, and confounding factors like class composition and time of year could not be controlled.

Future research can include longitudinal design following the same students from sixth through eighth grade to determine whether the shifts we observed occur within individual

students over time, or are a result of differences in the cohorts. Qualitative interviews could explore the reasoning between choices and survey answers. Including diverse socioeconomic and geographic populations would increase generalizability and potentially reveal differences. Most importantly, understanding these simultaneous shifts in girls and the divergence of well-being traits in boys can improve the design of well-being programs and collaboration in the classroom that work with students' natural development timelines.

Materials And Methods

Participants

*177 s*tudent volunteers in grade 6, 7 and 8 at Cresskill Middle School (Cresskill, New Jersey) participated in the study during the 2025-2026 school year. Participants included 65 sixth graders (36 girls, 29 boys) and 112 seventh and eighth graders (51 girls, 56 boys, with the remainder identifying as "prefer not to say" and excluded from gender-specific analysis) (Figure 1). Two teacher responses were identified and excluded from all analysis. Participation was voluntary. Written parental consent and minor assent were obtained for all participants prior to data collection. The study received approval with ISEF Human Participants Form 4 with full committee review, including signatures from a physician (MD), a science educator, and the school principal. The committee decided the study presented minimal risk and no protected groups involved. A qualified scientist (7th grade science teacher, MA+30, Middle School/Math/ Tech) directly supervised all procedures.

Survey instrument

Students completed a Google Form that included demographic information (grade and gender) and a Likert-scale well-being survey. Seven well-being traits were surveyed: health, optimism, resilience, belonging, trusted people, academic confidence, and hobby importance. Students also reported their work preference (prefer working alone vs. prefer working with others) on the same scale. All items used a 5-point Likert scale.

Behavior task

After the survey, students made an anonymous binary choice: share their answers to a series of puzzles with another student or complete the puzzles independently. To minimize the influence of social anxiety, pre-existing friendships, or classroom dynamics, sharing was digital and anonymous: students did not know who they might be paired with, and only the puzzle answers were shared (not the survey responses). After the choice was entered, the Google Form advanced to a series of puzzles. For participants who chose to share, a de-identified set of another student's puzzle answers appeared on their screen after they completed their own, and they were permitted to review and revise their answers. All participants, regardless of sharing choice, had to submit their answers within 15 minutes.

Standardization

The same script describing the study was used across all testing sessions. When the primary researcher could not be present due to scheduling conflicts, classroom teachers administered the experiment using the standardized scripts and protocol. Teachers posted the Google Form link via Google Classroom and students filled out the form on their school laptops.

Data collection and de-identification

Responses from the Google Form were collected in a Google Sheet. Data was deidentified before analysis to protect the privacy of the participants. No names or identifying information were listed on the data used for analysis.

Statistical analysis

Pearson correlations were calculated between each survey measure and the sharing decision (0=did not share, 1=shared) for the full sample and for subgroups by grade level (sixth grade vs seventh/eighth grade) and gender (girls vs. boys). Mean inter-trait r (MIr) was calculated using the absolute values of Pearson r coefficients to measure the strength of association vs. differentiation, despite direction. The direction of each trait pair was studied separately. Two sets of bar charts were created. The first showed the percentage of students who chose to share, broken down by grade level and gender (Figure 2). The second showed the mean difference in Likert scale scores between sharers and non-sharers for each survey trait, with 95% confidence interval bars, calculated separately for all participants and for subgroups by grade and gender (Figures 3-5). Fisher's r-to-z test was used to compare the strength of correlation between work preference and sharing across subgroups (Figure 6). Heat maps were created to display inter-trait correlation for each subgroup (Figure 7-8). Chi-square tests on raw sharing rates and a two-way ANOVA on the grade-by-gender interaction were also performed, but neither were statistically significant (chi-square $p=0.056$; ANOVA interaction $p>0.05$), suggesting that the correlation approach was more sensitive analysis tool for this study's questions.

Acknowledgements

We thank our Bergen SciChallenge mentor, Agatha Shaikhutdinova, for weekly virtual mentorship, statistical guidance, and presentation coaching through the project. We thank Alon Millet, for editing advice. We thank the sixth, seventh, and eighth grade teachers at Cresskill Middle School, specifically Ms. Sarah Travers, Ms. Suzanne Opfinger, Ms. Diana Zoino, Mr. Jason Timochko, Mr. Thomas Goeller, Ms. Leslie Janevski, and Ms. Elyssa Sawyer, who generously allowed us to conduct the study in their classrooms and followed our script when the primary researcher could not be present. We also thank all the students who volunteered to participate.

Works Cited

Ross, David A., et al. "Adolescent Well-Being: A Definition and Conceptual Framework." *Journal of Adolescent Health*, vol. 67, no. 4, 2020, pp. 472–476. https://doi.org/10.1016/j.jadohealth.2020.06.042

McLoughlin, Shane, and Kristján Kristjánsson. "Virtues as Protective Factors for Adolescent Mental Health." *Journal of Research on Adolescence*, vol. 35, 2025, e13004. https://doi.org/10.1111/jora.13004

Verma, Ritu. "Positive Psychology in Schools with Focus on Adolescent Well-Being." *Open Journal of Social Sciences*, vol. 12, no. 10, 2024, pp. 353–365. https://doi.org/10.4236/jss.2024.1210022

Waldinger, Robert J., and Marc S. Schulz. *The Good Life: Lessons from the World's Longest Scientific Study of Happiness*. Simon & Schuster, 2023.

Sandstrom, Gillian M., and Elizabeth W. Dunn. "Social Interactions and Well-Being: The Surprising Power of Weak Ties." *Personality and Social Psychology Bulletin*, vol. 40, no. 7, 2014, pp. 910–922. https://doi.org/10.1177/0146167214529799

Ascigil, Esra, et al. "Minimal Social Interactions and Life Satisfaction: The Role of Greeting, Thanking, and Conversing." *Social Psychological and Personality Science*, vol. 16, no. 2, 2025, pp. 202–213. https://doi.org/10.1177/19485506231209793

Gauvain, Mary, and Barbara Rogoff. "Collaborative Problem Solving and Children's Planning Skills." *Developmental Psychology*, vol. 25, no. 1, 1989, pp. 139–151. https://doi.org/10.1037/0012-1649.25.1.139

Swenson, Lisa M., and JoNell Strough. "Adolescents' Collaboration in the Classroom: Do Peer Relationships or Gender Matter?" *Psychology in the Schools*, vol. 45, no. 8, 2008, pp. 715–728. https://doi.org/10.1002/pits.20328