|CURIEUX|

ACADEMIC REVIEW

October Issue

Issue 12 Part 2

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: How Prejudice Became Law: Chinese Exclusion in Nineteenth-Century America By Aiden Wu

Page 24: The Consequences of Excessive Use of Digital Media from Early Childhood on the Adult Brain By Goomaral Sandag-ochir

Page 34: Computational Modeling, Performance Analysis, and Optimization of Wave and Tidal Energy Conversion Systems By Ishaan Srivastava

Page 49: Did the Expansion of Telemedicine in India (2020–2023) Correlate With a Reduction in Non-urgent Outpatient Department (OPD) Visits at Public Hospitals? By Sarvika Chaudhary

Page 61: From Separation to Synergy: Taking Advantage of Chemical and Biological Catalysts for Sustainable Synthesis By Rina Urabe

Page 76: An Anatomy of Allegiance: The Dual Nature of Fidelity in the East By Hrishik Malhotra

Page 83: Research Analysis on the Relationship Between Formal and Informal Education and the Income of Youth from SHE Clan Families in Tianchang City By Garrett Fulin Zhao

Page 91: From Sequence to Function: Programmable Structural Proteins as Sustainable Alternatives to Conventional Plastics By Lian Huan

Page 108: Evaluating Hand-Tracking Failures in Sign-Language Videos: The Roles of Skin Tone and Video Conditions By Kshiteej Herode

Page 118: Middle East Underdeveloped? A long-duration Exploration of Ottoman Decline and the Limits of Modernisation By Mir Hamid

Page 125: Nixon's Obsession With Image Control: The Watergate Scandal By Victoria Altirs

Page 138: Intermittent Fasting for Adolescent Athletes: A Literature Review By Yahya Dar

Page 145: How the Stages of Adolescence Affect Personality Traits By Mariela Vargas

Page 149: Neurological Symptom Burden and Repetitive Head Impact Exposure in Figure Skaters: A Cross-Sectional Survey By Amal Israilova

Page 163: The Impact of Quantum Computing on Classical Cryptography: Breaking RSA and ECC with Shor's and Grover's Algorithms By Sanjith Kota

Page 173: How Plants Can Be Grown In Space By Krishi Boppana

Page 176: ChatGPT and Social Etiquette By Nicole Wang

Page 183: Designing the Ethical Being: Moral Enhancement, Autonomy, and the Limits of Irrevocable Consent By Jamie Lee

Page 191: Recovering the In-Between: Retrofitting Back Lanes and Pocket Gardens in Mong Kok, Hong Kong By Gunho Jacob Ahn

Page 213: Improving Snake Robots Terrain Capability with a Passive Gripping Mechanism By Naman Gujral

Page 220: Can Music Reduce Stress in Adolescents? A Systematic Review of Music Listening and Cortisol Regulation By Prathiksai Amathi

Page 228: From Sharks to Humans: Comparative Tooth Regeneration Across Vertebrates and Their Implications for Human Regenerative Dentistry By Kate Kim

Page 243: From Arrival to Integration: An Analysis of Political Trust and Civic Engagement Among Afghan, Ukrainian, and Venezuelan Refugees in the United States By Long Nguyen

Page 264: Safe Space By Aanya Yadawad

Page 270: Too Much, Too Soon: Cultural Lag and Adolescent Exposure to Digital Pornography By Avery Solomon

Page 278: Marksman: An AI Powered Web Application for Intelligence Monitoring People and Objects By Anish Sharma

Page 289: The Effect of Low-Dose, Repeated Insect Repellent Exposure on Thermal Bleaching in Porites *cylindrica* By Issalyne Ladha

Page 299: Genetic and Biological Basis of Lean Type 2 Diabetes in South Asian Populations: A Narrative Review By Adithi Vattem

Page 308: Hormonal Physiology and Performance: How Menstrual-Cycle Tracking Improves Training and Reduces Injury in Female Athletes By Sai Pragna Prakash[1,2], Sai Pranati Prakash[3]

Page 325: FBN-1 Driven Phenotypic Changes in Cardiac Tissue and the Aortic Wall in Patients with Marfan Syndrome By Tanvi Nulkar

Page 340: The Role of LSM14B in Eukaryote Oocyte Development and Cancer By Anvita Lakhotia

Page 353: When Tourism Becomes “Too much”: Resident-perceived Overtourism Thresholds in Kyoto and Tokyo By Koichiro Shindo

Page 379: Stem Cell-Based Therapeutics in Amyotrophic Lateral Sclerosis: Therapeutic Potential and Safety Challenges By Koji Karasawa

Page 407: Luxury Logos and Social Identity: How Designer Brands Shape Status Perception By Mogoș Maria-Alexandra

Page 419: Evaluating Early-Career BMI Trajectory for Predicting NBA Career Longevity Using Machine Learning By Darsh Goel

Page 432: From the Course to the Simulator: Off-Course Golf and the Technology-enabled Experience EconomyBy Le Tuan Minh

Page 440: The Dark Side of the Pygmalion Effect By Vaishali Sita Noronha

Page 450: Unraveling the Problem: How Fabric Design Can Help End Microplastic Pollution By Savera Dhaliwal

Page 460: Reassessing the Safe-Haven Premium: A Regression-Based Evaluation of U.S. Treasury Convenience Yield Underperformance Relative to Stock–Bond Correlation By Taimur Lodhi

Page 480: How Dance Enhances Memory: What Neuroscience Reveals About Movement, Music, and the Aging Brain By Zhirui Liu

Page 487: The Economic Effects of Emerging E-Commerce Platforms on U.S. Retail Competition By Trevor Wu

Page 501: The Anatomy of Dismissal: Ancient Bodies, Ancient Control Gender Discrimination in Women's Medical Treatments Across the Classical World By Manuela Mejía-Fanjul

Page 510: Polarizability of PEG-600 through VQE By Keerthan Challa

Page 517: Breast Cancer Awareness: What Are the Things Young Girls and Women Can Do Every Day to Prevent Breast Cancer? By Adora Liao

Page 523: The Aftermath of The 2008 Economic Crisis Regarding Firms, Households, and Financial Institutions By Dan Vu

Page 532: A Lost Page in History: Identifying the First Asian American US Senate Page By Jayden Kim, Advisor By Katherine A. Scott, Ph.D.

Page 551: microRNAs in Alzheimer's Disease: Amyloid -β and Tau Regulation and Their Usage as Biomarkers and Therapeutic Targets By Khurshid Jimmy Hormazdiyar

Page 562: Chemical Strategies for Mitigating Jeju Island's Groundwater Contamination By Hyunsung Ko

Page 573: Housing Stability and Access to Opportunity: A Comparative Study of Low-Income Housing Policy and Urban Planning in New York City, Hong Kong, and Tokyo By Zhixiao Max Liu

Page 584: Mental Toughness in Adolescents By Olivia Shekou

Page 593: Treatment of Alzheimer's disease using Stem Cells and CRISPR-based genome editing systems By Teruki Otake

Page 612: Bridging the Adherence Gap: The Role of MCT Supplementation and Exogenous Ketones in Pediatric Epilepsy Management By Mustafa Tornaci

Page 630: Quantitative EEG Analysis of Brain Activity Before, During, and After Nadi Shodhana Pranayama By Maya Kumari

Page 647: How Can Poison Become Medicine? Medicine and Pharmaceutical Processing in Medieval China through the Rise and Fall of Five-Stone Powder (五石散) By Wushi San

How Prejudice Became Law: Chinese Exclusion in Nineteenth-Century America
By Aiden Wu

Keywords

Immigration, Chinese Exclusion, Prejudice, Legislation, Stereotypes, Culture

Abstract

This paper examines the Chinese Exclusion Act as a manifestation of evolving public perceptions regarding Chinese immigrants during the late nineteenth century. It posits that the Act was not merely a product of nativist sentiment or economic competition, but was shaped by complex cultural stereotypes and legislation that systematically othered Chinese communities. Focusing on political cartoons, congressional testimony, Supreme Court rulings, and treaty and statutory developments from the Burlingame–Seward Treaty (1868) through the Chinese Exclusion Act (1882) and its legislative extensions, the paper traces how economic anxieties and racialized stereotypes were legitimized into national policy. Using analysis of cartoons by Thomas Nast and G. F. Keller, alongside testimony from the 1877 Joint Special Committee, congressional debate, and landmark Supreme Court decisions, the paper argues that depictions of Chinese people as unassimilable provided the moral and political rationale for exclusion. The analysis mobilizes theoretical frameworks from Elizabeth Young-Bruehl (psychic roots of prejudice), Mary Douglas (purity/pollution), and René Girard (scapegoating) to show how political art reflected and reinforced these xenophobic tropes. The paper draws parallels between anti-Chinese imagery and rhetoric and the resulting exclusionary legal actions, highlighting the legacy of those frameworks in shaping Chinese Americans' role in society.

The Chinese Exclusion Act, passed on May 6, 1882, was one of the first major federal immigration laws to bar entry by a specific national-origin group. It marked a turning point because racial restriction became an open principle of national policy. Chinese immigrants were not the only group to face discrimination; Irish, Italian, and other immigrant communities also encountered hostility. However, Chinese migrants, especially laborers, were singled out at a time when race, rights, and national belonging were being reshaped through policy and societal forces.

Beginning in the late 1860s, anti-Chinese hostility circulated through public debate, political cartoons, and local politics. This paper argues that economic anxiety, racial prejudice, visual culture, and legal institutions worked together to transform local resentment into federal action. Over time, these forces made restrictions appear lawful and compatible with the nation's stated ideals about labor, liberty, and democracy.

The first major wave of Chinese immigration began when around 25,000 Chinese migrants, primarily men, landed in California in 1851 and were welcomed because labor was in high demand in the American West. In 1868, Congress ratified the Burlingame-Seward Treaty, which supported relatively open migration between China and the United States (Office of the Historian, n.d.). However, this openness soon gave way to economic tensions. Many of these

laborers were hired at lower wages by companies seeking to cut costs. As the job market tightened, white workers increasingly blamed Chinese workers for falling wages and rising unemployment, especially in railroad construction, mining, and garment production (DeSimone, 2022).

Labor unions and politicians capitalized on these fears, portraying Chinese workers as both economic threats and unassimilable outsiders. Over time, these accusations escalated into racial hostility and violence, including the Los Angeles Chinese massacre of 1871, during which nineteen Chinese immigrants were murdered (Wallace, 2017). By the late 1870s, politicians and industrial leaders began to treat them not simply as labor competitors but as a danger to public order and morality. Reflecting these concerns, Congress created the Joint Special Committee to Investigate Chinese Immigration in 1877. Testimony before the Committee revealed how economic concerns, cultural anxieties, and racial hierarchy merged into a broader argument for exclusion. Frank M. Pixley, former attorney general of California, exemplified this logic when he stated:

"The burden of our accusation against them is that they come in conflict with our labor interests; that they can never assimilate with us; that they are a perpetual, unchanging, and unchangeable alien element that can never become homogeneous; . . . that they can never become citizens, and that an alien, degraded labor class, without desire of citizenship, without education, and without interest in the country it inhabits, is an element both demoralizing and dangerous to the community within which it exists" (United States Congress, 31).

This testimony shows how economic grievances became tied to a racial claim about permanent foreignness. By framing the Chinese as both workers and racial outsiders, witnesses like Pixley helped turn prejudice into political evidence. In 1882, Congress passed the Chinese Exclusion Act, barring most Chinese immigration.

The rest of this paper traces that process in three stages. Part I analyzes how media portrayals gradually spread the view that Chinese immigrants were unassimilable and threatening. Part II explores how sentiment became federal policy through the Page Act, the Joint Special Committee investigation, the Angell Treaty, and finally the Chinese Exclusion Act. Part III examines how Supreme Court rulings narrowed immigrants' rights and gave exclusion constitutional legitimacy. Together, these sections show how anti-Chinese prejudice moved from popular hostility to legislation to judicial approval, becoming embedded in national law.

Part I: Cartoons as Political Argument

In *Anatomy of Prejudices*, Elizabeth Young-Bruehl explains that prejudice works by constructing a social group as "the Other." Racism, in particular, turns visible differences such as hair texture, skin color, or physical features into signs of inferiority. She also argues that racial prejudice can project fears and anxieties onto an outgroup (Young-Bruehl, 108–115). In nineteenth-century anti-Chinese imagery, political art often portrayed Chinese figures as

"virtually inhuman, monstrous, or non-existent" (Young-Bruehl, 133), encouraging readers to view them as people to fear, mock, or exclude.

Other scholars help explain why these depictions carried such political force. Mary Douglas's definition of "dirt" as "matter out of place" clarifies how racist imagery presented Chinese immigrants as bodies that did not belong within the social order (Douglas, 3). René Girard's theory of scapegoating further explains how social tension could be redirected toward a vulnerable minority, allowing the majority to unite against a common target (Girard, 16). These ideas help explain how cartoons transformed prejudice into political arguments that made exclusion appear justified. The following images show how anti-Chinese cartoons developed this logic in visual form.

An Inclusive Civic Ideal

Figure 1: Uncle Sam's Thanksgiving Dinner, 20 November 1869, by Thomas Nast (Nast, 1869)

Figure 1, by Thomas Nast, depicts people from different racial and ethnic groups seated together at a Thanksgiving dinner. The caricatures are relatively subdued, and the guests are portrayed with dignity. The characters interact warmly, creating a sense of welcome. Columbia, a female symbol of America, faces a Chinese guest seated with his wife and child. Uncle Sam stands at the head of the table. Behind him, a painting labeled "Welcome" shows Castle Garden, an immigration center in New York. These details present an inclusive national ideal in which Chinese immigrants appear capable of joining the American community. Later cartoons, however, would reverse this image and instead portray them as outsiders to be rejected. This progression resulted from a variety of reasons, including the economic panic of 1873 where white workers scapegoated Chinese immigrants for lower wages and lack of jobs. This xenophobia was amplified by opportunistic politicians, labor leaders, and other immigrant groups leading to a profound shift from careful acceptance to hostility.

Visual Normalization of Exclusion

Figure 2: Uncle Sam's Thanksgiving, by G. F. Keller (Keller, 1877)

By 1877, anti-Chinese political cartoons increasingly used national symbols to mark Chinese immigrants as outsiders. A representative example is George Frederick Keller's *Uncle Sam's Thanksgiving*, published by *The Wasp*, a satirical magazine based in San Francisco, where such hostility was especially intense.

In this image (figure 2), several racial and ethnic figures surround the table, reflecting Keller's view of the U.S. racial hierarchy. The white Uncle Sam sits confidently at the head. A Black man appears as a servant presenting the turkey. A Native American crouches on a stool, eating with his hands, drawing on primitivist stereotypes. European figures are also caricatured, but still participate more fully in the meal, suggesting that they remain assimilable.

Most notably, the Chinese man is depicted as alien and radically different. His face is exaggerated, he is dressed as a foreigner, and he eats a rat with chopsticks while a white man laughs. This image marks an escalation in racist imagery. Keller does not simply show him as unfamiliar; he turns him into a figure of disgust. Through Young-Bruehl's idea of projection (Young-Bruehl, 34), the cartoon can be read as placing fear and disorder onto the Chinese figure by making him seem almost inhuman. The rat-eating scene is particularly disturbing because it gives this fear a visual form. At a table meant to symbolize the nation, he is shown as foreign, unclean, and dangerous. This also fits Mary Douglas's definition of "dirt" as "matter out of place" (Douglas, 3): he is treated as someone whose presence violates the moral boundary of society.

Ironically, Keller himself was an immigrant from Prussia. However, his use of this cartoon exemplifies a view that white European immigrants were still different but assimilable. In contrast to Nast's 1869 piece, this image portrayed Chinese people as immoral and incapable of civic life. It shows how the perception of the Chinese immigrant shifted from a possible member of the national community to a figure whose exclusion seemed natural.

The Denial of Entry

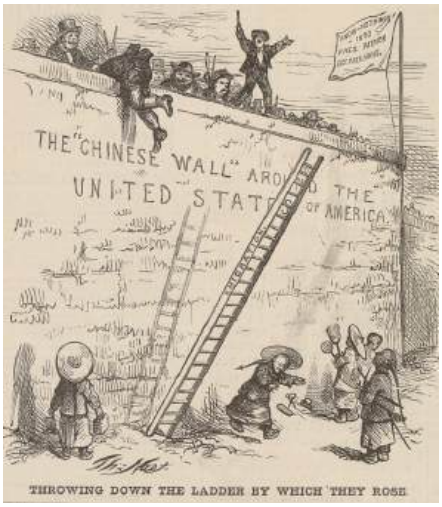


Figure 3: Throwing Down the Ladder by Which They Rose, by Thomas Nast (Nast, 1870)

This image by Thomas Nast(figure 3) depicts European immigrants climbing the ladder of "emigration," while Chinese migrants remain at the bottom. The "Chinese Wall" symbolizes a barrier to entry. As the last immigrant reaches the top of the wall, he kicks the ladder away while another celebrates, suggesting that America is open to Europeans but closed to the Chinese. A Know Nothing flag waves in the scene, recalling an anti-immigration movement originally formed to oppose Irish and Catholic immigrants decades earlier.

The cartoon reflects the belief that European immigrants could rise into American society while the Chinese could not. Nast captures the hypocrisy of some immigrant groups: once accepted themselves, they could deny the same opportunity to others. Girard's idea of scapegoating helps explain this pattern: Chinese immigrants became a "common enemy" whose exclusion allowed other groups to strengthen their own claim to acceptance (Girard, 16).

In this climate, some immigrant groups aligned themselves with mainstream white working-class opinion by casting the Chinese as more foreign than themselves. This pattern is particularly clear among Irish immigrants, who had faced nativist hostility but later became some of the loudest voices against the Chinese (McCarron, 8). In this way, the cartoon reveals a racial hierarchy that would later appear in law: Europeans might be absorbed, while Chinese immigrants remained outside the nation's boundaries.

The Conflict over National Identity

Figure 4: “The Chinese Question” by Thomas Nast (Nast, 1871)

“The Chinese Question”(Figure 4) captures a conflict between competing visions of the nation. At the center stands Columbia, the allegorical female embodiment of the United States. Her upright posture, calm expression, and white dress symbolize moral authority and democratic ideals. She positions herself between a Chinese man and a hostile white mob. He appears wounded and vulnerable, emphasizing his powerlessness. Yet, he is still caricatured, marking him as different even in a cartoon that defends him.

Behind them, the writing on the wall includes racial slurs such as “coolie,” “slave,” “rat-eater,” “heathen,” and “barbarian.” Above Columbia, it states, “I (a white man) am opposed to him on the grounds of 1. Race, 2. Industry, 3. Politics, 4. Morality.” The first-person phrasing, “a white man,” reveals how whiteness functioned as a political status. By listing race, labor, politics, and morality together, the wall turns separate anxieties into one argument for exclusion.

Nast’s cartoon brings together the patterns seen in Figures 1–3. The Chinese figure is portrayed not simply as a worker but as a threat to social order itself. The wall turns fear into accusation, marks him as outside the community, and makes him a convenient target for blame. In that way, the cartoon combines projection, boundary-making, and scapegoating. Columbia represents the conflict between inclusive constitutional promises and exclusionary nationalism. However, the crowd’s size and aggressiveness show how easily public opinion could overpower democratic principles. The wall’s accusations anticipate arguments that later appear in hearings, statutes, and court opinions. By the late 1870s, the inclusive vision shown in Nast’s 1869 cartoon had been largely replaced by images portraying Chinese residents as foreign, dangerous, and removable**.**

Part II: Anti-Chinese Hostility in Law

The Chinese Exclusion Act was one of the most consequential immigration laws in nineteenth-century American history. It did not appear suddenly. Its passage emerged through treaties, investigations, and statutes that gradually narrowed the rights of Chinese migrants (“Chy Lung v. Freeman: Anti-Chinese Sentiment and the Supremacy of Federal Immigration Law”).

Through the Burlingame-Seward Treaty, the Page Act, the Joint Special Committee investigation, the Angell Treaty, and the Chinese Exclusion Act, anti-Chinese hostility moved from public debate into federal immigration law.

The Burlingame-Seward Treaty of 1868: Partial Inclusion

The Burlingame-Seward Treaty grew out of expanding trade and diplomatic relations with Qing China alongside increasing Western dependence on Chinese migrant labor. The treaty affirmed broad principles of mobility and reciprocal protection, including the rights of travel, residence, education, and religious freedom. Through this language, the treaty portrayed Chinese migrants as lawful and entitled to legal protection.

However, the treaty's openness had limits. Article VI made clear that these rights did not extend to naturalization (Burlingame-Seward Treaty). It therefore created a distinction between economic acceptance and political status, allowing Chinese workers to be useful as laborers while still being denied citizenship. As economic competition intensified in the 1870s, nativist rhetoric began to cast Chinese workers as threats to both the economy and culture (Choy, Dong, and Hom, 48). Many white laborers, particularly in California, argued that Chinese workers undercut wages by accepting lower pay and harsher conditions. Following the Panic of 1873, anti-Chinese sentiment accelerated, framing them as "servile tools of monopolistic corporations" and enemies of the white working class (Rosalsky). Nativist critics also claimed that Chinese migrants lived apart from white communities, maintained distinct customs, and resisted assimilation. Anti-Chinese activists further accused them of bringing vice, gambling, prostitution, and opium use into American cities (Ahmad, 2000). These claims expanded labor competition into a broader attack on Chinese morality, sexuality, and social order.

The Page Act of 1875: Moral Regulation

The Page Act shows how debates about Chinese labor also became arguments about morality, gender, and race. Often identified as one of the first restrictive federal immigration laws, the Act targeted Asian migrants under the name of barring forced laborers and women entering for "lewd and immoral purposes" (Page Act). Although its language applied broadly, it was enforced especially against Chinese women. Immigration officials were given the power to determine whether an Asian woman was being trafficked for immoral purposes, allowing them to discriminate against and harass Chinese women ("Anti-Asian laws and policies").

The Page Act had two important effects. First, it reduced the number of Chinese women entering the United States; from 1880 to 1882, over 50,000 Chinese men reportedly entered the United States while only about 550 women were admitted (Women & the American Story "Page Act"). Second, it established a precedent by which immigration law could be leveraged as a tool for moral regulation and racialized restriction, turning stereotypes about Chinese women's immorality into legal barriers.

The Joint Special Committee Investigation: Federal Legitimation

By the mid-1870s, public hostility toward Chinese laborers pressured Congress to authorize a Joint Special Committee to investigate Chinese immigration. Its hearings in 1876 and report in 1877 marked an important shift: anti-Chinese prejudice entered the official records. The investigation brought these arguments into the national political debate.

A recurring question in the hearings was whether Chinese immigrants could assimilate. Frank M. Pixley, former attorney general of California, exemplified this concern. When asked whether they could "become attached to our institutions and reconcile to live and die here," Pixley responded that "They have no love for our institutions. They have no knowledge of our institutions." He further argued that they "take no step in the direction of citizenship" and "acquire no homes." Through this language, cultural stereotypes became political evidence that Chinese immigrants could not form lasting ties to the country (United States Congress, 16-17).

Pixley also framed labor competition as a threat to the moral and familial foundations of white working-class life. When asked what jobs Chinese laborers occupied, he stated:

"Laundry-work, cigar-making, slippers, sewing-machine labor, they have very nearly monopolized. They are largely employed as domestic servants and as office-boys. . . . [I]n silk and woolen manufactories, in fruit-picking, in gardening, in harvesting, . . . in railroad-building, in placer-mining, . . . in fishing and peddling fish, are among the most noted of their industries, and from these industries that I have named they have nearly driven out the entire white labor" (United States Congress, 17).

This statement reflects a common anti-Chinese narrative that portrays Chinese workers as an invasive force displacing white workers. His claims that they "very nearly monopolized" certain industries and had "nearly driven out the entire white labor" recast economic competition as a threat to white labor and social order.

Pixley extended this argument when asked, "How does their employment affect our white labor?" He contrasted the "true American hero," a married white worker who supported his wife and children, with the Chinese worker, whom he described as "an adult male who has no wife, no family, no child." This stereotype ignored the fact that policies such as the Page Act had severely restricted Chinese women's immigration. Pixley thus treated Chinese bachelor communities as evidence of cultural deficiency rather than a product of law.

Other officials, including George H. Rogers and Frank McCoppin, members of the California State Senate, echoed similar claims that Chinese immigrants were incompatible with American institutions and cultural norms. Testimonies from figures such as Henry H. Ellis, the San Francisco chief of police, and Dr. Meares, head of the city Board of Health, reinforced similar arguments through claims about public safety, morality, and health. The multitude of different claims reflects how the council was more focused on finding reasons for exclusion rather than fair evaluation of the Chinese community.

The Angell Treaty of 1880: Treaty Revision

By 1880, political pressure pushed the federal government to renegotiate the Burlingame-Seward Treaty. On November 17, 1880, the United States and China signed the

Angell Treaty amid growing pressure over Chinese immigration. The treaty stated that it was responding to the "constantly increasing immigration of Chinese laborers to the territory of the United States, and the embarrassments consequent upon such immigration" (Angell Treaty, 1). By describing immigration itself as an "embarrassment," the treaty framed Chinese laborers as a social and political burden.

The treaty's key language appeared in Article I:

"Whenever in the opinion of the Government of the United States, the coming of Chinese laborers to the United States, or their residence therein, affects or threatens to affect the interests of that country, or to endanger the good order of the said country or of any locality within the territory thereof, the Government of China agrees that the Government of the United States may regulate, limit, or suspend such coming or residence, but may not absolutely prohibit it" (Angell Treaty, 1).

This provision gave the United States authority to "regulate, limit, or suspend" the immigration of Chinese laborers whenever the government claimed that their presence threatened national interests or public order. Although the treaty stated that immigration could not be "absolutely prohibited[ed]," it made Chinese migration a conditional privilege subject to federal control. This marked a major departure from the Burlingame-Seward Treaty's earlier promise of mutual free migration. By authorizing restriction, the Angell Treaty gave anti-Chinese sentiment a formal diplomatic basis and made full exclusion increasingly likely.

The Chinese Exclusion Act of 1882: Final Codification

Just two years after the Angell Treaty, Congress used this newly established authority to pass the Chinese Exclusion Act. Signed by President Chester A. Arthur in 1882, the Act suspended the immigration of Chinese laborers for ten years and required exempt travelers, such as merchants, students, and diplomats, to obtain certification from the Chinese government. Echoing the treaty almost verbatim, it justified exclusion by asserting that "the coming of Chinese laborers… endangers the good order of certain localities," turning a vague concern about public order into federal law (Chinese Exclusion Act).

The Act also imposed documentation requirements that made entry, reentry, and lawful residence increasingly conditional. Chinese individuals had to obtain official certification verifying their identity and eligibility, placing their movement under a new system of state scrutiny. Through these requirements, the Act reframed Chinese migrants not as participants in mutual migration but as people to be monitored, restricted, and expelled.

Many of the same grievances against Chinese immigrants surfaced in congressional debates leading up to the Act. In his address to the Senate, Senator John P. Jones of Nevada made this argument especially clear:

"The people in every country in Europe have struggled for liberty on every battle-field from Marston Moor to Magenta. They have our instincts; they have our tastes. Their God is our God. They are inherently lovers of liberty. . . . When and where did the Chinese race ever make a single declaration for liberty? When and where did they strike one single blow for freedom or

free institutions? . . . [T]he United States is the last and best fruit of a struggle that has been waged by the Caucasian race for thirty centuries in favor of freedom, in which the Chinese have borne no part. Every fiber in their heart, every corpuscle in their blood has been molded in the spirit of despotism under which they have lived uncomplainingly" (United States Congress, 1741).

This rhetoric framed exclusion as a question of racial fitness for democracy. It drew a sharp hierarchy: Europeans were portrayed as natural heirs of liberty while Chinese immigrants were depicted as incapable of democratic life.

Jones continued this racial logic by presenting prejudiced generalizations about Chinese laborers:

"You cannot take some exceptional individual and judge his whole race by him. You must take into consideration the history and characteristics of a whole race. . . . What, then, I ask, has been the contribution from China? Oppression, barbarism, degradation. A civilization purely material, nothing spiritual about it; everything commutable in money; no injury that a man can inflict on a Chinaman that he is not willing to take a money compensation for; no sensibility to be ruffled, no honor to be outraged. I do not speak of individuals, but I speak of the race" (United States Congress, 1741).

Rather than acknowledging individual variation, Jones argued that entire races should be judged by collective "history and characteristics." This framing presented prejudice as objective analysis rather than as bias. Chinese people were portrayed not merely as different but as lacking basic moral and emotional qualities. By claiming to speak not of individuals but of "the race," Jones suggested that no personal meritor or exception could overcome racial judgment.

Some senators spoke in opposition to the law. Senator Joseph E. Brown of Georgia, for example, defended Chinese immigrants in limited terms and warned Congress not to treat them as a monolith:

"Before the vote is taken I simply wish to remark that I do not desire to see the naturalization of Chinese in this country, but I am not satisfied that under the treaties as they now exist . . . we have a right to refuse to naturalize them while we naturalize other foreigners. There are some Chinese who are now here who are intelligent people, and some of them are assimilating with us, some of them are members of churches, some of them are Christian men. There are some of them who would make good citizens. Therefore, I should be unwilling to say, without further investigation upon that point, that there was no case where a Chinaman could properly become a citizen of the United States. . . . For the present bill, and in the present state of things under the treaties, I am not prepared to vote for the entire exclusion of Chinese from naturalization" (United States Congress, 1747).

Brown's statement shows that opposition to exclusion existed, but it was heavily qualified. He challenged total exclusion more than the racial assumptions behind it, and he openly conceded that he did not wish to see Chinese naturalization broadly.

Taken together, the laws, investigations, and debates preceding the Chinese Exclusion Act did more than amplify anti-Chinese sentiment; they gave it official form. The 1882 Act also

became the foundation for later restrictions. In 1888, the Scott Act barred the reentry of Chinese laborers who had temporarily left the United States, stranding thousands whose return had previously been permitted. In 1892, the Geary Act extended the ban for another ten years and required Chinese residents to carry certificates of residence at all times while also curbing their civil rights. Not until the Magnuson Act of 1943 were the exclusion laws formally repealed; even then, Congress permitted only 105 Chinese immigrants per year to enter and become eligible for naturalization, preserving the restrictive spirit of the earlier system (Magnuson Act). The Act was therefore not a temporary reaction but the beginning of a long federal structure of restriction.

Part III: Supreme Court Rulings

The restrictions described in Part II had major consequences for Chinese immigrants and, eventually, Chinese Americans communities. According to census data (Cheng, 2023), the Chinese population grew rapidly from roughly 4,000 in 1850 to 64,000 in 1870, but increased only slightly from approximately 105,000 in 1880 to 107,000 in 1890, showing the effects of new limits on settlement and family formation. The Supreme Court also played a major role in defining the legal boundaries of this exclusion. Through its decisions, the Court recognized certain basic rights while granting the government wide authority over entry, reentry, and expulsion. *Chew Heong v. United States* and *Yick Wo v. Hopkins* offered important but narrow protections. *Chae Chan Ping v. United States* and *Fong Yue Ting v. United States* then gave much broader force to federal power over exclusion and removal. Finally, Justice Harlan's dissent in *Plessy v. Ferguson* revealed how normalized anti-Chinese exclusion had become: even a celebrated defense of Black equality relied on Chinese degradation as a baseline. These cases show how the Court turned the assumptions described in Parts I and II into constitutional doctrine.

Chew Heong v. United States *and* Yick Wo v. Hopkins*: Narrow Protection*

One early landmark decision came in *Chew Heong v. United States* (1884). As discussed in Part II, the 1882 Act suspended the immigration of Chinese laborers and required those leaving the country to obtain certificates proving their right to return. Chew Heong, a Chinese laborer who had lived in the United States before the Act, left in 1881 before the requirement applied to him. When he attempted to return, he was denied entry because he did not possess one.

After being detained in San Francisco, Heong challenged his exclusion, arguing that the certificate requirement could not fairly apply to someone who had left before such documents were available. He relied on rights protected by the 1880 Angell Treaty, which protected certain returning Chinese laborers (Price, "Chew Heong v. United States"). The Supreme Court, in a majority opinion by Justice Harlan, ruled for Heong, holding that Congress had not clearly intended to take away treaty-protected return rights for laborers in his position.

However, the ruling was not unanimous. Justice Field, joined in part by Justice Bradley, dissented on the ground that the statute's plain language required the certificate for return and that Congress could supersede treaty obligations through later legislation. More tellingly, Bradley's dissent reflected a broader distrust of Chinese immigrants. He dismissed the lower

courts' acceptance of testimonial evidence by claiming it "was calculated to produce great abuses, for Chinese of the lower class have little regard for the solemnity of an oath." His assertion that Congress could override treaty obligations revealed how fragile Chinese immigrants' rights remained (*Chew Heong v. United States*).

A second important victory came in *Yick Wo v. Hopkins* (1886). This case arose from a San Francisco ordinance requiring laundry operators to obtain permits from the Board of Supervisors if their businesses were housed in wooden buildings. On its face, the law appeared to be a neutral public safety regulation, but its enforcement told a different story. Almost all of the more than two hundred Chinese applicants were denied permits while nearly all non-Chinese applicants operating under similar conditions were approved. Yick Wo, a Chinese laundry operator who had run his business lawfully for decades, was arrested and fined for operating without a permit. His case eventually reached the Supreme Court.

The Court ruled unanimously in his favor, holding that the ordinance had been "applied and administered by public authority with an evil eye and an unequal hand." This established the principle that discriminatory enforcement of a neutral law was itself unconstitutional. The Court also held that the Fourteenth Amendment's Equal Protection Clause extended to "all persons within the territorial jurisdiction" of the United States, regardless of race or citizenship. The decision affirmed that Chinese immigrants, despite being barred from naturalization and targeted by exclusionary federal law, were still entitled to basic constitutional protections in daily life (*Yick Wo v. Hopkins*).

Chew Heong and *Yick Wo* stand as rare victories for Chinese immigrants, but their protections were limited. *Chew Heong* protected treaty-based return rights, while *Yick Wo* affirmed that the Equal Protection Clause applied to all persons within U.S. jurisdiction. Neither decision, however, challenged Congress's authority to restrict Chinese immigration. Their importance, therefore, lies not only in what they protected, but also in what they left untouched.
Chae Chan Ping v. United States*: Sovereign Exclusion*

The limits of these protections became clearest in *Chae Chan Ping v. United States* (1889). Chae Chan Ping was a Chinese laborer who had lived in the country for over a decade and held a valid certificate guaranteeing his right to leave and return. While he was abroad, however, Congress passed the Scott Act of 1888, which invalidated the certificates and barred the reentry of Chinese laborers, including those who had legally resided in the United States and complied with earlier requirements. Similar to the *Chew Heong* case, Chae Chan Ping was denied entry despite possessing documentation that had been valid when he left. Unlike *Chew Heong*, however, the Court allowed Congress to invalidate that earlier protection.

The Supreme Court upheld his exclusion and issued one of its strongest statements about federal immigration power. Writing for the Court, Justice Stephen J. Field framed the issue not as one of individual rights, but as one of national sovereignty. His opinion declared that "the power of the legislative department…to exclude aliens…is an incident of sovereignty" essential to independence and self-preservation. This reasoning allowed the Court to sidestep the conflict between the Scott Act and earlier treaties with China, even while acknowledging that the statute

was “in contravention of express stipulations” of those agreements. Instead, the Court held that treaties and federal statutes were equal under the Constitution, meaning that “the last expression of the sovereign must be controlled” (*Chae Chan Ping v. United States*). In practice, this allowed Congress to override earlier treaty commitments.

The Court treated exclusion as a political question largely beyond judicial review, emphasizing that Congress could exclude aliens “whenever the public interests require it, without judicial interference.” It also distinguished between vested property rights, which might survive the replacement of a treaty, and conditional privileges, such as a Chinese laborer’s permission to return. In the Court’s view, whatever permission Chinese laborers had previously received was “held at the will of the government, revocable at any time at its pleasure” (*Chae Chan Ping v. United States*).

The decision also repeated familiar racial and civilizational arguments for exclusion. The Court depicted Chinese laborers as economically threatening, culturally unassimilable, and socially destabilizing, suggesting that their continued presence could endanger “the peace of the community” and even “the preservation of our civilization” (*Chae Chan Ping v. United States*). By using this language, the Court recast exclusion as national self-defense rather than racial restriction.

This ruling marked a major departure from the more protective stance in *Chew Heong* and established what later became known as the “plenary power” doctrine: the principle that Congress holds sweeping authority over immigration, largely free from judicial constraint. By allowing Congress to erase earlier protections in the name of national self-preservation, the Court placed Chinese laborers in a deeply vulnerable position: present enough to be regulated but still foreign enough to be excluded.

Fong Yue Ting v. United States*: Deportation and Surveillance*

The Supreme Court extended this logic four years later in *Fong Yue Ting v. United States* (1893). The case arose under the Geary Act of 1892, which continued the Chinese exclusion regime and required all Chinese laborers to carry a certificate of residence. If found without one, they were subject to arrest and removal unless they could prove their lawful presence through the testimony of “at least one credible white witness.” Fong Yue Ting, along with two other Chinese laborers, was detained despite having legally resided in the United States since before 1882. Without a “credible white witness,” they faced expulsion.

Fong Yue Ting extended the logic of *Chae Chan Ping* from exclusion at the border to removal within the country. The Court asserted that the power to expel foreigners was “as absolute and unqualified as the right to prohibit and prevent their entrance.” It grounded this authority in national self-preservation and cited the principle that a nation may bar those who threaten “the peace of the State” or “corrupt the manners of the citizens.” This reasoning carried plenary power from exclusion to deportation. The “credible white witness” requirement made this racial hierarchy explicit: legal and moral credibility was defined through whiteness. As a

result, Chinese residents were subject to surveillance, documentation, and removal in ways that European immigrants generally were not (*Fong Yue Ting v. United States*).

Harlan's Dissent in **Plessy v. Ferguson*****: Exclusion as Legal Common Sense***

The Supreme Court's influence on Chinese immigrants extended beyond immigration law. In *Plessy v. Ferguson* (1896), which upheld Louisiana's Separate Car Act and established the "separate but equal" doctrine, Justice Harlan's famous dissent still relied on some of the same racial logic used against Chinese migrants. As the only dissenter from the pro-segregation ruling, Harlan argued that the majority's decision was not only unjust but also dangerous. Because the destinies of white and Black Americans were "indissolubly linked together," he reasoned, the state had an interest in preventing, not deepening, racial hostility. Laws like Louisiana's, which assumed that "colored citizens are so inferior and degraded that they cannot be allowed to sit in public coaches occupied by white citizens," would cause social division by planting "the seeds of race hate under the sanction of law." For Harlan, legal segregation would produce racial conflict rather than peace between the races (*Plessy v. Ferguson*, 45–46).

However, Harlan's defense of Black equality also depended on a comparison that reinforced rather than challenged Chinese inequality. He stated:

"There is a race so different from our own that we do not permit those belonging to it to become citizens of the United States. Persons belonging to it are, with few exceptions, absolutely excluded from our country. I allude to the Chinese race. But, by the statute in question, a Chinaman can ride in the same passenger coach with white citizens of the United States, while citizens of the black race in Louisiana … are yet declared to be criminals … if they ride in a public coach occupied by citizens of the white race" (Plessy v. Ferguson, 47).

Rather than rejecting racial hierarchy as a whole, Harlan used Chinese exclusion as a foil to expose the injustice of Louisiana's Separate Car Act. He pointed out that Chinese people were considered so "different" that the United States denied them citizenship and had "absolutely excluded" them from the country. Yet, he noted with irony, a Chinese person could ride in the same passenger coach as white citizens, while Black citizens could be punished for doing the same.

In effect, Harlan used the degraded legal status of Chinese people as a baseline against which Black citizenship could be defended. His rhetoric reveals that anti-Chinese sentiment was so normalized by 1896 that even one of the Court's strongest critics of segregation could treat Chinese exclusion as a rhetorical floor rather than an injustice in its own. Harlan's opinion shows how deeply anti-Chinese assumptions had entered constitutional argument even outside immigration law.

Overall, the decisions analyzed in Part III reveal a consistent pattern: the Supreme Court alternated between narrow protection and broad deference to exclusionary power. The Court used the language of sovereignty, civilization, and public order to justify that authority. As a result, Chinese people could live under U.S. law, but their place within it remained conditional. This conditional status became one of the most important legal legacies of Chinese exclusion.

Part IV: Conclusion

Chinese exclusion was part of a longer history of racialized immigration control. Labor conflict and racial prejudice mattered, but the broader pattern was more systematic. Political cartoons, congressional investigations, federal statutes, treaty revisions, government enforcement, and Supreme Court decisions together turned anti-Chinese hostility into a national system of restriction. From the earliest years of the United States, access to citizenship had been racially limited by the Naturalization Act of 1790, which restricted naturalization to "free white persons." Over the next century, that racial logic extended into immigration law.

The American nation of this period defined itself partly by deciding who could not fully enter it. Chinese immigrants served this role: they became the outsider against whom whiteness, citizenship, and national identity could be defined. The Page Act of 1875, which sharply restricted Chinese women's immigration under the presumption of immorality, produced the very conditions later used to justify exclusion. Many Chinese men were prevented from forming families, Chinese women were treated as threats to public morality, and the law itself hardened these stereotypes. The Supreme Court then gave this system constitutional force. Although *Chew Heong v. United States* and *Yick Wo v. Hopkins* offered narrow protections, *Chae Chan Ping v. United States* and *Fong Yue Ting v. United States* treated exclusion and deportation as matters of national sovereignty largely insulated from ordinary judicial review.

This history also changes how we understand post-Reconstruction America. Reconstruction is usually framed around Black freedom and its violent suppression, and rightly so. However, at the same time, the federal government was building another system of racial control, one focused on entry, residence, and legal status. Chinese immigrants were central to that development. To ignore this history is not simply to leave a gap in the record; it is to misunderstand how the United States became the country it did. Chinese exclusion was therefore not just one discriminatory episode. It became one of the major ways the United States defined who could enter, who could remain, and who could be treated as permanently foreign even after years of living under American law.

Works Cited

Ahmad, Diana L. "Opium Smoking, Anti-Chinese Attitudes, and the American Medical Community, 1850–1890." American Nineteenth Century History, vol. 1, no. 1, 2000, doi:10.1080/14664650008567016.

Chae Chan Ping v. United States, 130 U.S. 581 (1889). Immigration History, immigrationhistory.org/item/chae-chan-ping-v-united-states-1889-aka-the-chinese-exclusion-case/.

Cheng, Jerry. "Chinese-Americans 1785– : Demographics." SCSU Research Guides, Southern Connecticut State University, 2023, libguides.southernct.edu/c.php?g=15048&p=81577.

Chew Heong v. United States, 112 U.S. 536 (1884). Justia, supreme.justia.com/cases/federal/us/112/536/.

Choy, Philip P., et al. Coming Man: 19th Century American Perceptions of the Chinese. University of Washington Press, 1995.

DeSimone, Barbara. "The Chinese Exclusion Act, Part 1 – The History." In Custodia Legis, Library of Congress, 13 May 2022, blogs.loc.gov/law/2022/05/the-chinese-exclusion-act-part-1-the-history/.

Douglas, Mary. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Routledge & Kegan Paul, 1966.

Evans, Charles, editor. Treaties and Other International Agreements of the United States of America 1776–1949. Vol. 6, Library of Congress, tile.loc.gov/storage-services/service/ll/lltreaties/lltreaties-ustbv006/lltreaties-ustbv006.pdf

Federal Judicial Center. "Chy Lung v. Freeman: Anti-Chinese Sentiment and the Supremacy of Federal Immigration Law." Federal Judicial Center, fjc.gov/history/spotlight-judicial-history/chinese-immigration-restriction.

Fong Yue Ting v. United States, 149 U.S. 698 (1893). Justia, supreme.justia.com/cases/federal/us/149/698/.

"Fong Yue Ting v. United States (1893)." Immigration History, immigrationhistory.org/item/fong-yue-ting-challenge-case/.

Girard, René. The Scapegoat. Translated by Yvonne Freccero, Johns Hopkins University Press, 1986.

Keller, George Frederick. Uncle Sam's Thanksgiving. San Francisco Wasp, 1877. Thomas Nast Cartoons, thomasnastcartoons.com/2014/11/23/americas-thanksgiving-celebration-1869/keller-thanksgiving.

Mao, K. J. "On the Mere Presumption: The Page Act of 1875 and the Ramifications of Racialized Immigration Policy on Chinese American Women." Hunter College High School Research Archive, 2024, rais.education/wp-content/uploads/2024/08/0428.pdf.

McCarron, Bryan P. The Global Irish and Chinese: Migration, Exclusion, and Foreign Relations Among Empires, 1784–1904. 2016. Georgetown University, PhD dissertation.

Georgetown University Digital Repository, repository.digital.georgetown.edu/handle/10822/1040727.
Nast, Thomas. Uncle Sam's Thanksgiving Dinner. Harper's Weekly, 20 Nov. 1869, thomasnastcartoons.com/2014/11/23/americas-thanksgiving-celebration-1869/.
Nast, Thomas. Throwing Down the Ladder by Which They Rose. Harper's Weekly, 23 July 1870, thomasnastcartoons.com/2015/01/06/throwing-down-the-ladder-by-which-they-rose-23-july-1870/.
Nast, Thomas. The Chinese Question. Harper's Weekly, 18 Feb. 1871.
https://thomasnastcartoons.com/2016/02/13/the-chinese-question/
National Archives. "Chinese Exclusion Act (1882)." National Archives, 2023, archives.gov/milestone-documents/chinese-exclusion-act.
National Park Service. "Anti-Asian Laws and Policies." National Park Service, 2025, nps.gov/articles/000/anti-asian-laws-and-policies.htm.
Office of the Historian. "The Burlingame-Seward Treaty, 1868." U.S. Department of State, history.state.gov/milestones/1866-1898/burlingame-seward-treaty.
Plessy v. Ferguson, 163 U.S. 537 (1896). Legal Information Institute, Cornell Law School, law.cornell.edu/supremecourt/text/163/537.
Price, Anna. "Chew Heong v. United States." In Custodia Legis, Library of Congress, 26 Jan. 2023, blogs.loc.gov/law/2023/01/chew-heong-v-united-states/.
Rosalsky, Greg. "The Price America Paid for Its First Big Immigration Crackdown." Planet Money, NPR, 26 Nov. 2024, npr.org/sections/planet-money/2024/11/26/g-s1-35805/chinese-expulsion-act-railroads-immigration-crackdown.
United States, Congress. Additional Articles to the Treaty between the United States and China of June 18, 1858 (Burlingame-Seward Treaty). 1868, govinfo.gov/content/pkg/STATUTE-16/pdf/STATUTE-16-Pg739.pdf.
United States, Congress. The Page Act of 1875. 1875, loveman.sdsu.edu/docs/1875Immigration%20Act.pdf.
United States, Congress. Angell Treaty. 1880, govinfo.gov/content/pkg/STATUTE-22/pdf/STATUTE-22-Pg826.pdf#page=1.
United States, Congress, Joint Special Committee to Investigate Chinese Immigration. Report of the Joint Special Committee to Investigate Chinese Immigration. Government Printing Office, 1877, babel.hathitrust.org/cgi/ssd?id=coo.31924073426482.
United States, Congress. Chinese Exclusion Act. 1882, loveman.sdsu.edu/docs/1882ChineseExclusionAct.pdf.
United States, Congress. Congressional Record. 47th Cong., 1st sess., 9 Mar. 1882, p. 1741.
United States, Congress. Congressional Record. 47th Cong., 1st sess., 9 Mar. 1882, p. 1747.
United States, Congress. Magnuson Act (Chinese Exclusion Repeal Act). Public Law 78-199, 57 Stat. 600, 17 Dec. 1943.

Wallace, Kelly. "Forgotten Los Angeles History: The Chinese Massacre of 1871." Los Angeles Public Library Blog, 19 May 2017, lapl.org/collections-resources/blogs/lapl/chinese-massacre-1871.
Women & the American Story. "Page Act, 1875." New-York Historical Society, 2026, wams.nyhistory.org/industry-and-empire/expansion-and-empire/page-act-1875/.
Yick Wo v. Hopkins, 118 U.S. 356 (1886). Justia, supreme.justia.com/cases/federal/us/118/356/.
Young-Bruehl, Elisabeth. The Anatomy of Prejudices. Harvard University Press, 1996.

The Consequences of Excessive Use of Digital Media from Early Childhood on the Adult Brain By Goomaral Sandag-ochir

Abstract

As the world becomes increasingly digital, technology and social media have emerged as significant factors in everyday life, influencing cognitive, emotional, and behavioural development. This study investigates the neurobiological risks of excessive digital media use among children and adolescents, with an emphasis on critical neurodevelopmental processes such as synaptogenesis, myelination, and synaptic pruning. Long-term exposure to screens during vital periods of brain maturation has been associated with alterations in both grey and white matter, especially in regions implicated in decision-making, reward processing, and executive function. Evaluative studies conducted by Lee et al. (2019), Small et al. (2020), and Horvath et al. (2020) demonstrate that problematic smartphone use correlates with decreases in grey matter volume in the orbitofrontal cortex and structural changes within the anterior cingulate cortex. These abnormalities are linked to heightened vulnerability to addiction, diminished self-regulation, and reduced attention span. Moreover, prolonged exposure to digital interactions stimulates dopamine release, which may contribute to the development of addictive behaviours, alter neural connectivity, and impair cognitive efficiency. Collectively, these findings illuminate the long-term neurological and psychological ramifications of extensive technology exposure in early developmental stages. This research underscores the need to understand the effects of digital exposure on neurodevelopmental health and cognitive functioning in younger populations, advocating for regulatory measures and increased awareness regarding digital media usage.

Keywords digital media, neurodevelopment, cognitive function, adolescent behaviour

1. Introduction

Go watch TV. A commonly said phrase by parents to keep their children distracted. Living in a modern digital age, digital media has become a ubiquitous part of daily life for many individuals, ranging from children, adolescents and adults. The beloved social media platforms used by billions of individuals across the world facilitate effortless access to information, sharing of ideas and thoughts through virtual networks and communities, promoting the ability to connect with others to make the world feel more *interconnected*. From Facebook, Instagram, TikTok, YouTube, Twitter, and countless other applications, social media algorithms are designed to promote user engagement by delivering endless personalised content through user interactions and behaviours to ensure prolonged usage and addiction (Kulke). With its compelling nature, social media has been found to display problematic smartphone user behaviours and fail to regulate addiction despite the obvious negative psychosocial consequences of anxiety, FOMO, and depression (Liu et al., 2024). However, little to no information about the neurobiological mechanisms is recognised. As more and more younger children engage with

digital devices, long-term effects on neurodevelopment can endanger the key opportunities for growth towards the adult brain.

One question that should be examined is - how many hours of social media per day is regarded as excessive? Multiple experts, such as Hunt and Young (2018) and Withington and Punch (2019), recommended that 30 minutes per day can lead to significant improvement in mental health. While 2 hours per day on average is the limit that adolescents and adults should aim towards. Additionally, the AACAP recommends that children have a screen time of an hour per day (American Academy of Child and Adolescent Psychiatry, 2020). While the use of more than 3 hours of social media per day can increase the risks of mental health problems in adolescents by more than 60% (Riehm et al.). In addition to mental health disorders, a study by eClinicalMedicine found that among more than 10,000 adolescents who used excessive social media, correlation with decreased sleep cycles, which are associated with lack of concentration, memory loss and poor academic performance (Paakkari et al.). These, however, can be adjusted for by several external variables like sleep quality, physical activity levels, and socioeconomic status. Adolescents with fewer physical outlets or from lower SES levels might experience more screen exposure and disrupted circadian rhythms, amplifying digital media's cognitive and affective effects.

In the first few years of a child's life, the central nervous system undergoes significant neurological changes and shows great cerebral plasticity, which is crucial for optimal neurodevelopment to adolescence and early adulthood (Centre on the Developing Child). During this critical neurological period, there are high levels of neural connections (synaptogenesis) and dendritic growth, processes of myelination, and synaptic pruning, which shape the neural structure for cognitive and behavioural functioning (Lemaître et al.). These processes then highly influence the maturation of neural circuits, and cognitive and social behaviour as a child matures.

Myelination is a crucial adaptive process that provides the building foundations for brain connectivity by accelerating the electrical transmissions between neurons by supporting cognitive and behavioural functioning (Susuki, 2010). This process follows a distinct temporal pattern that starts from the primitive structures of brainstem regions shown in *Figure 1.1,* while moving upwards from the back of the cerebral cortex to the last brain region called the prefrontal cortex, which is responsible for higher cognitive functions, ultimately reconfiguring the brain connectivity into adulthood.

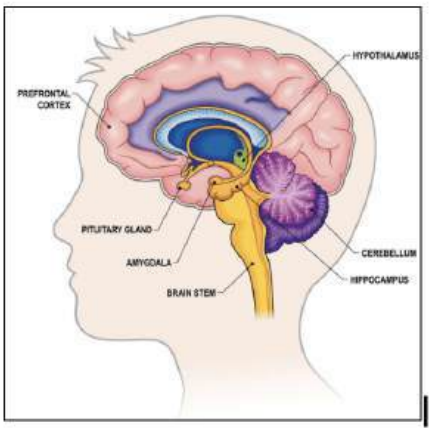


Figure 1.1: *'The anatomy of the human brain'*

Synaptic pruning is the process by which the brain removes neurons and synapses to ensure maximum efficiency in brain function. Synaptic maturation usually reaches its peak at the age of 2 or 3, which then reduces as an individual develops, as shown in *Figure 1.2* (Huttenlocher). As the brain contains millions of neurons, which communicate with one another using a combination of electrical signals, they are then converted into chemical signals carried by molecules called *neurotransmitters* (National Library of Medicine). During the earliest stages of an embryo's development, simpler associations are replaced by more complex structures, allowing the removal of unwanted synapses and neuronal structures as the human brain develops (Rowden). This, in turn, increases the efficiency of the brain's neural transmissions, conserving energy, streamlining neural connections, by enhancing and creating space for critical neurotransmitters.

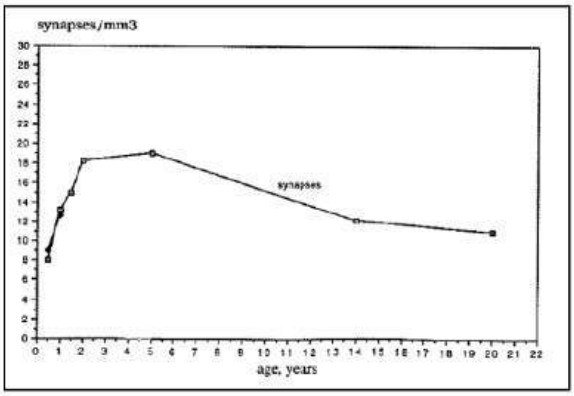


Figure 1.2: *'Estimation of total synapse number in human visual cortex'*

White and grey matter are two key components of the central nervous system in the brain and spinal cord, as shown in *Figure 1.3,* which can have significant impacts on cognitive and psychic functions. Grey matter consists of the unmyelinated neurons, synapses, and glial cells, in which processing of information through signalling is found in white matter functions (Physiopedia). The key areas of grey matter include the cerebral cortex, basal ganglia, limbic system, and the spinal cord. These properties enable individuals to control motor skills,

behavioural and emotional responses, memory, and send motor commands (Mercadante and Tadi). Whilst white matter found in subcortical regions of the brain, contains myelinated bundles of axons wrapped with myelin (a substance of proteins and lipids), which helps and increase the nerve signals (Villines). There are three different types of fibres within the white matter region, including projection, association, and commissural, as shown in *Figure 1.4.* These connections then facilitate the transmission of electrical signals, supporting the body to process information. Throughout developmental periods from the ages of 8, grey matter volume increases. The density of grey matter increases until the age of 20 from hormonal changes from puberty, and synaptogenesis, which allows higher processing and further mental development (Mercadante and Tadi). Conversely, white matter keeps developing well into 20 years of age due to its complex neural networks, and processes of myelination and synaptic pruning within the region.

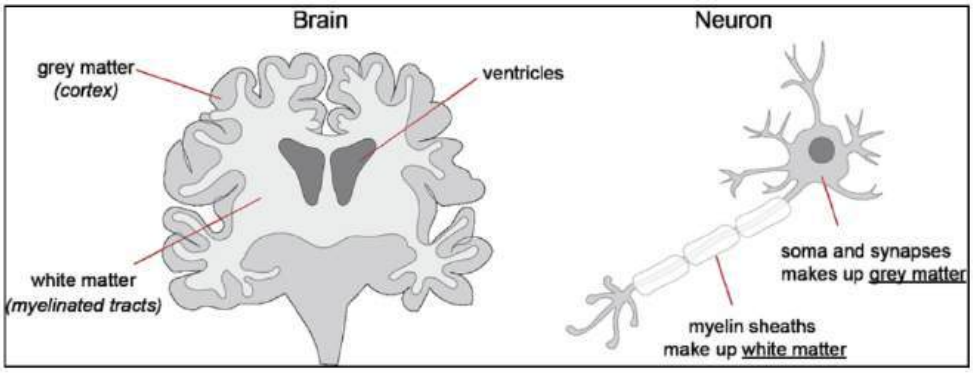


Figure 1.3: *'White and Gray Matter'* (Villines, 2015).

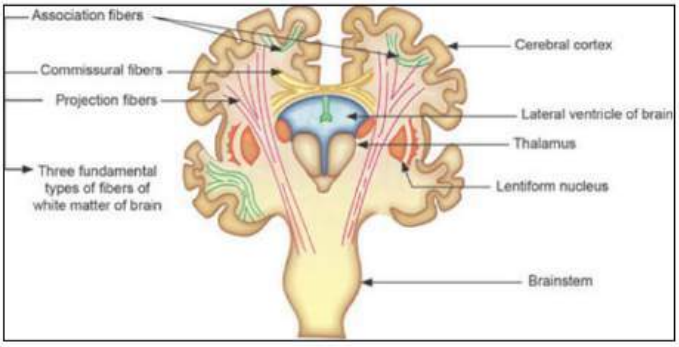


Figure 1.4: *'White Matter regions'*

Several neuroscientific studies have accumulated empirical evidence about the maturation and neurodevelopmental processes that occur in the human brain during the first years of life. However, with the increase of technology concerns regarding social skills, attention, and sleep disruption, it has been linked with neurodevelopment throughout the years. Researchers, including Lee et al. (2019), Small et al. (2020), Hutton et al. (2020) and Meri et al. (2023), focus on how the excessive use of digital media can have negative effects on the neurodevelopment of children and adolescents into adulthood.

2. The addiction behind digital and social media

Dopamine, a G-protein coupled receptor, is a crucial neurotransmitter that regulates our perception of pleasure, motivation and reward as stated in research conducted by Horvath et al (2020). Dopamine release upon activating the reward system through activities like scrolling, playing video games or watching videos gives immediate satisfaction and feel-good chemicals. At times, this pleasant sensation may make individuals pursue it again, causing possible addiction. The structure and function of the brain can be altered by sustained exposure to digital rewards with frequent dopamine release that leads to adjustments in connection strengths between neurons. Individuals with smartphone addictions had lower grey matter volumes in their insula and temporal cortex, along with reduced resting-state activity of the anterior cingulate cortex, according to Horvath et al. (2020). The production of dopamine involves changing the tyrosine amino acid into L-DOPA, which is then made into enzymes that eventually become dopamine. Dopamine is managed by the hypothalamus situated within the brain, as shown in *Figure 2.1*. However, when there is not enough dopamine, one might develop Parkinson's disease, schizophrenia, depression, lack of motivation and concentration problems (Watson). Therefore, maintaining a balance in dopamine levels is crucial. Some actions that can be taken to balance dopaminergic neurons are to reduce screen time, consumption of unhealthy food, alcohol, drugs, smoking, and other activities that release high levels of dopamine.

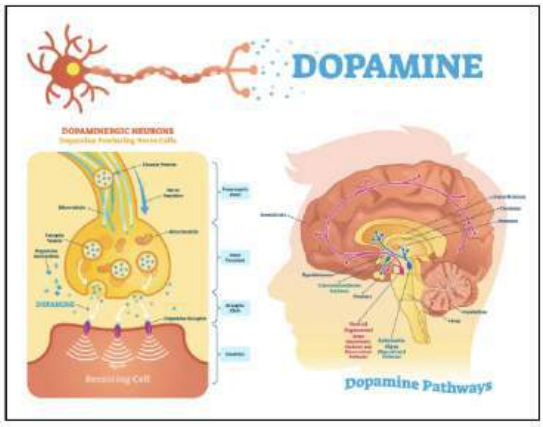


Figure 2.1: '*Dopamine Pathways*'

3. Brain health consequences of digital technology use

From the research conducted at the Department of Psychiatry and Biobehavioral Sciences and Semel Institute for Neuroscience and Human Behaviour, doctors of medicine Gary W.Small et al. (2020) found that frequent digital technology use has a significant impact, both negative and positive, on brain function and behaviour. By assessing the neural activation in 24 middle-aged and older adults, it was found that 12 of the subjects had minimal experience with the internet (labelled as net-naive), and the other 12 had extensive experience (labelled as net-savvy). The study found that text reading activated brain regions involved in language, reading, memory, and visual processing in both net-naive and net-savvy groups. However, during internet searching, the net-savvy group showed significantly more extensive activation in brain

regions controlling decision-making, complex reasoning, and vision compared to the net-naive group (Small et al.). An MRI, as shown in *Figure 2.2,* was used to record brain neural activity before and after internet training. The testing is presumed for a week, with an hour per day consisting of brief instructions on how to search online with practice sessions. As shown from the scans, during the first session, the net-naive group showed brain activity in specific parts of the *'superior, middle, and inferior frontal gyri, as well as the lateral occipital cortex and occipital pole'* (Small et al.). During the second session, they started using more parts of their frontal gyri, hippocampus, entorhinal cortex, and anterior cingulate, as demonstrated in *Figure 2.3.* This shows how the mental activities supporting tasks required for internet searches, including motivation, decision-making, working memory, and the ability to suppress irrelevant information, significantly increased. However,r for the first session, the net-savvy group used more of their frontal, temporal and optical lobe. In contrast to the second session, *Figure 2.2* shows that brain activity in those specific areas significantly decreased. This reduction displays consistently how the brain becomes more efficient and possibly habituates to certain tasks individuals can easily complete over time.

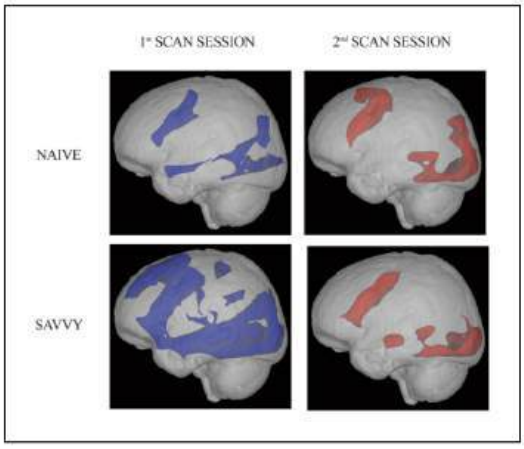


Figure 2.2: *'MRI scans before and after internet sessions'*

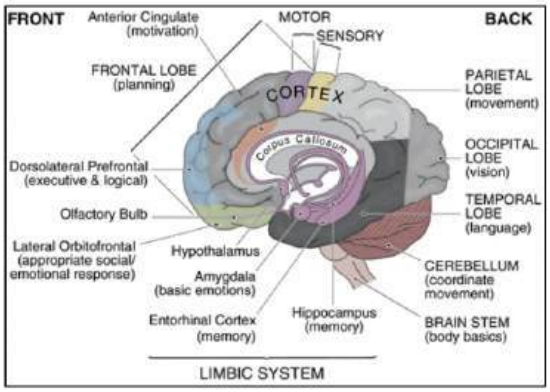


Figure 2.3: *'Diagram of human brain properties'*

4. Lateral orbitofrontal grey matter abnormalities in subjects with problematic smartphone use

The study 'Lateral orbitofrontal grey matter abnormalities in subjects with problematic smartphone use' by Lee et al. (2019) explores the fronto-cingulate grey matter abnormalities among those who overuse social networking platforms. The study included 39 individuals with problematic smartphone addictions and 49 normal control male and female smartphone users. A Voxel-based morphometric analysis (VBM) based on diffeomorphic anatomical registration with an exponentiated Lie algebra algorithm to examine regional disparities between groups in terms of grey matter volume (GMV) and fronto-cingulate region (Lee et al.). From the results, there was a reduction in GMV in the right lateral orbitofrontal cortex (OFC) among problematic smartphone users compared to the healthy controls (Lee et al.). Furthermore, strong negative correlations were observed between GMV in the right lateral OFC and the Smartphone Addiction Proneness Scale (SAPS) score, including the SAPS tolerance subscale (Lee et al.). As Small GMV within the lateral OFC showed a tendency towards higher immersion on smartphones for long durations, these results suggest that grey matter alterations may be associated with problematic use of smartphones, especially through overuse of digital media. Similarly ,a study by Hutton et al. (2020) demonstrates that *'human-technology interactions may be linked to the grey matter composition of brain regions which govern behaviour'.* Therefore, these findings indicate that abnormality within the orbitofrontal cortex exerts a greater impact upon regulatory control over previously reinforced behaviours and may underlie the problematic digital media use.

The potential harmful effects of extensive technology use are found to heighten attention-deficit disorders, abnormalities in brain function, and influence cognitive and behavioural functioning. The investigation aimed to determine the neurobiological risks of excessive use of digital media from early childhood towards the adult brain. Therefore, the research reveals the need for caregivers to pay close attention to early neurodevelopment, as it can have a powerful impact on a child's cognitive, emotional, and social development, as well as their physical and mental health.

Works Cited

American Academy of Child and Adolescent Psychiatry (2020). Screen Time and Children. [online] www.aacap.org. Available at: https://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/Children-And-Watching-TV-054.aspx#:~:text=For%20children%202%2D5%2C%20limit [Accessed 16 Jun. 2024].

Brain Waves (2016). The Brain - Diagram and Explanation. [online] Brainwaves.com. Available at: http://www.brainwaves.com/ [Accessed 18 Jun. 2024].

Cataldo, I. (2021). Social Media Usage and Development of Psychiatric Disorders in Childhood and Adolescence: a Review. [online] Frontiers in Psychiatry. Available at: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.508595/full#F2 [Accessed 25 Jun. 2024].

Center on the Developing Child (2007). InBrief: The Science of Early Childhood Development. [online] Center on the Developing Child at Harvard University. Available at: https://developingchild.harvard.edu/resources/inbrief-science-of-ecd/#:~:text=Brains%20are%20built%20over%20time%2C%20from%20the%20bottom%20up.&text=In%20the%20first%20few%20years [Accessed 10 Jun. 2024].

Cleveland Clinic (2022). Dopamine. [online] Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/articles/22581-dopamine [Accessed 20 Jun. 2024].

Deb, S. (2014). JaypeeDigital | Cerebrum— White Matter. [online] www.jaypeedigital.com. Available at: https://www.jaypeedigital.com/book/9789350909409/chapter/ch9 [Accessed 10 Jun. 2024].

Guy-Evans, O. (2022). The Role of Dopamine as a Neurotransmitter in the Human Brain - Simply Psychology. [online] Simply Psychology. Available at: https://www.simplypsychology.org/the-role-of-dopamine-as-a-neurotransmitter-in-the-human-brain.html [Accessed 20 Jun. 2024].

Horvath, J., Mundinger, C., Schmitgen, M.M., Wolf, N.D., Sambataro, F., Hirjak, D., Kubera, K.M., Koenig, J. and Christian Wolf, R. (2020). Structural and functional correlates of smartphone addiction. Addictive Behaviors, 105, p.106334. doi:https://doi.org/10.1016/j.addbeh.2020.106334.

Hunt, M. and Young, J. (2018). (PDF) No More FOMO: Limiting Social Media Decreases Loneliness and Depression. [online] ResearchGate. Available at: https://www.researchgate.net/publication/328838624_No_More_FOMO_Limiting_Social_Media_Decreases_Loneliness_and_Depression [Accessed 2 Jun. 2024].

Huttenlocher (1990). 13-Synaptic Density and Estimation of Total Synapse Number in Human Visual Cortex. [online] Research Gate. Available at: https://www.researchgate.net/figure/Synaptic-density-and-estimation-of-total-synapse-number-in-human-visual-cortex-From_fig20_265536911 [Accessed 10 Jun. 2024].

Hutton, J.S., Dudley, J., Horowitz-Kraus, T., DeWitt, T. and Holland, S.K. (2020). Associations Between Screen-Based Media Use and Brain White Matter Integrity in Preschool-Aged

Children. JAMA Pediatrics, [online] 174(1), p.e193869. doi:https://doi.org/10.1001/jamapediatrics.2019.3869.

Hutton, J.S., Jessica Taylor Piotrowski, Bagot, K., Blumberg, F., Turhan Canli, Chein, J., Christakis, D.A., Grafman, J., Griffin, J.A., Hummer, T., Kuss, D.J., Lerner, M., Marcovitch, S., Paulus, M.P., Perlman, G., Romeo, R., Thomason, M.E., Turel, O., Weinstein, A. and West, G. (2024). Digital Media and Developing Brains: Concerns and Opportunities. Current Addiction Reports, 11. doi:https://doi.org/10.1007/s40429-024-00545-3.

Kulke, S. (2023). Social media algorithms exploit how we learn from our peers. [online] news.northwestern.edu. Available at: https://news.northwestern.edu/stories/2023/08/social-media-algorithms-exploit-how-humans-learn-from-their-peers/ [Accessed 2 Jun. 2024].

Lee, D., Namkoong, K., Lee, J., Lee, B.O. and Jung, Y.-C. (2019). Lateral orbitofrontal gray matter abnormalities in subjects with problematic smartphone use. Journal of Behavioral Addictions, 8(3), pp.404–411. doi:https://doi.org/10.1556/2006.8.2019.50.

Lemaître, H., Augé, P., Saitovitch, A., Vinçon-Leite, A., Tacchella, J.-M., Fillon, L., Calmon, R., Dangouloff-Ros, V., Lévy, R., Grévent, D., Brunelle, F., Boddaert, N. and Zilbovicius, M. (2020). Rest Functional Brain Maturation during the First Year of Life. Cerebral Cortex (New York, NY), [online] 31(3), pp.1776–1785. doi:https://doi.org/10.1093/cercor/bhaa325.

Liu, N., Zhu, S., Zhang, W., Sun, Y. and Zhang, X. (2024). The relationship between fear of missing out and mobile phone addiction among college students: the mediating role of depression and the moderating role of loneliness. Frontiers in public health, [online] 12. doi:https://doi.org/10.3389/fpubh.2024.1374522.

Mercadante, A.A. and Tadi, P. (2023). Neuroanatomy, Gray Matter. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK553239/#:~:text=The%20grey%20matter%20has%20a [Accessed 16 Jun. 2024].

National Library of Medicine (1997). The Principles of Nerve Cell Communication. Alcohol Health and Research World, [online] 21(2), pp.107–108. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6826821/#:~:text=Two%20mechanisms%20have%20evolved%20to [Accessed 18 Jun. 2024].

Paakkari, L., Tynjälä, J., Lahti, H., Ojala, K. and Lyyra, N. (2021). Problematic Social Media Use and Health among Adolescents. International Journal of Environmental Research and Public Health, [online] 18(4). doi:https://doi.org/10.3390/ijerph18041885.

Physiopedia (2024). Grey and White Matter. [online] Physiopedia. Available at: https://www.physio-pedia.com/Grey_and_White_Matter#:~:text=Grey%20matter%20volume%20increases%20until,past%2020%20years%20of%20age. [Accessed 20 Jun. 2024].

Rowden, A. (2023). What is synaptic pruning? [online] www.medicalnewstoday.com. Available at:

https://www.medicalnewstoday.com/articles/synaptic-pruning#:~:text=The%20neurons%20in%20the%20brain [Accessed 12 Jun. 2024].

Small, G. (2020). Brain health consequences of digital technology use. Dialogues in Clinical Neuroscience, [online] 22(2), pp.179–187. doi:https://doi.org/10.31887/dcns.2020.22.2/gsmall.

Susuki, K. (2010). Myelin: A Specialized Membrane for Cell Communication. [online] Nature.com. Available at: https://www.nature.com/scitable/topicpage/myelin-a-specialized-membrane-for-cell-communication-14367205/ [Accessed 26 May 2024].

The College of Arts and Sciences (2023). Study shows habitual checking of social media may impact young adolescents' brain development | UNC-Chapel Hill. [online] The University of North Carolina at Chapel Hill. Available at: https://www.unc.edu/posts/2023/01/03/study-shows-habitual-checking-of-social-media-may-impact-young-adolescents-brain-development/#:~:text=The%20study%20findings%20suggest%20that [Accessed 20 Jun. 2024].

Villines, Z. (2015). Gray Matter vs. White Matter in the Brain. [online] Spinalcord.com. Available at: https://www.spinalcord.com/blog/gray-matter-vs-white-matter-in-the-brain [Accessed 21 Jun. 2024].

Watson, S. (2021). Dopamine: The pathway to pleasure. [online] Harvard Health. Available at: https://www.health.harvard.edu/mind-and-mood/dopamine-the-pathway-to-pleasure#:~:text=Neurons%20in%20the%20region%20at [Accessed 20 Jun. 2024].

Withington, S. and Punch, A. (2019). There are Costs from Spending Too Much Time on Social Media. [online] Maxwell School. Available at: https://www.maxwell.syr.edu/research/lerner-center/population-health-research-brief-series/article/there-are-costs-from-spending-too-much-time-on-social-media [Accessed 21 May 2024].

Computational Modeling, Performance Analysis, and Optimization of Wave and Tidal Energy Conversion Systems By Ishaan Srivastava

Abstract

Marine Energy is one of the most underused renewable forms of energy on the planet. Despite the potential it has, the deployment and implementation of marine energy systems have been slow. This is due to the engineering limitations, environmental concerns and different challenges that occur per location. This study aims to look at how wave and tidal energy systems perform at different locations in the world, pulling data in from the System Advisor Model (SAM) to provide empirical evidence. The research question was as follows: Does accounting for site-specific conditions alongside technology efficiency lead to better and more efficient deployment strategies to improve system performance for marine energy? To test this idea, wave and tidal energy systems at multiple locations with different environmental wave conditions were looked at. Some factors assessed were resource availability, energy conversion efficiency, and cost of efficiency, and these factors showed the performance differences per different site locations. The results showed that the system output differed a lot depending on the local conditions per site, and as predicted, there were certain regions showing much stronger potential for the implementation of marine energy technologies. There were several optimization strategies in mind per site location, such as hybrid system integration and adaptive design modifications, which would be based on site-specific information and improved site selection. These approaches were able to strongly increase energy output while also decreasing the financial burden and improving marine energy technology's ease of use. All in all, the findings in this study suggest that tailoring marine energy technologies to specific site locations can improve the effectiveness and allow for increased production for the future.

Introduction

Renewable energy is defined as energy gained from sources that can replenish their supply at a higher rate than they are consumed. Fossil fuels, for example, classified as non-renewable energy, take hundreds of millions of years to form and produce harmful greenhouse gases when burned to produce energy. Unlike this, renewable energy sources are plentiful and naturally found all around us in nature, making them a sustainable and environmentally friendly alternative to power homes, businesses, and transportation, all examples of technology used by humans constantly. The common sources of renewable energy are solar, which captures sunlight; wind, which uses the movement of the air to generate electricity; geothermal, which captures heat from the earth; hydropower, which converts the energy of flowing freshwater bodies; and ocean energy, which captures energy from the ocean.

Ocean energy harnesses the energy from the ocean and converts it into usable energy, and there are two types: wave and tidal energy. For wave energy, there are mainly three types of technology used the Wave Buoy, Wave Oscillating Surge Converter (OSC), and Wave Oscillating

Water Column (OWC). For tidal energy, there are mainly two types of technology used, the Horizontal Turbine and the Vertical Turbine.

Starting with the wave energy systems, firstly, the point absorber, or the wave buoy, is a piece built around a buoy that sits on the water's surface and bobs up and down with passing waves. Beneath it, a pump and generator system converts that vertical motion into electricity. It can do this because as the buoy rises and falls, it pulls and pushes a rod that is connected to the pump, driving moving parts that spin a generator. The whole device is anchored to the seafloor by a mooring system that keeps it in position while still allowing vertical movement. The mechanical simplicity of the design, combined with its ability to operate in open water far from shore, makes this model one of the most widely developed and efficient wave energy technologies.

Secondly, the oscillating wave surge converter uses the horizontal push of waves near the seafloor, unlike the point absorber which uses the up-and-down motion at the surface. The main component of the device is a large paddle-like flap that is anchored to the seabed, and it is fixed at its base. The device is positioned in the nearshore zone where the wave surge is the strongest, which allows for the largest energy capture. As the waves pass, the pressure difference across the flap causes it to swing back and forth with the motion of the water. This back-and-forth motion drives a hydraulic piston, which pressurizes fluid through a circuit which then powers a hydraulic engine, which then generates electricity. The flap moves while the base stays fixed, and so the system is mechanically straightforward and simple, and so it is well suited for high energy coastal environments.

Lastly, the oscillating water column device captures wave energy through its partially underwater chamber which is then open to the sea at its base. As the waves enter this chamber, they push the internal water column upwards, compressing the air that is trapped above it. As the wave recedes, the water level drops and the air expands again, and then this cycle is repeated. This continuous compression and release drives air back and forth through a turbine mounted at the top of a chamber. Then that turbine is connected to a generator that produces electricity. OWC devices can be integrated into some coastal structures that already exist, such as breakwaters, which makes OWC devices a cost-effective option for shoreline installations where there is already infrastructure.

Now, talking about the tidal energy systems, the first is the horizontal-axis turbine tidal stream generator, which works on the same idea and technology as the wind turbine but is submerged in ocean water and is driven by tidal currents instead of air. The turbine blades are mounted horizontally on a shaft and are fixed to the seabed through a support tower, which is then positioned in channels or straits where currents run strongest. As water flows past, it pushes against the blades and causes them to rotate which in that spinning motion drives a generator that produces electricity. This electricity is then carried back to shore through underwater cables as these devices are usually placed far beyond the shoreline. Horizontal turbines are good at capturing energy of fast directional flows, and so they are the most common type deployed for tidal energy and are most proven at the commercial level.

Lastly, the vertical-axis turbine tidal steam generator works similarly to the horizontal-axis device, but with the blades arranged around a vertical shaft rather than a horizontal one. This orientation means that the turbine can capture the tidal flow from any direction, which gives this specific design an advantage in ocean locations where currents shift or are less consistent (11). As the water moves past the blades, they rotate around the central axis, turning the shaft and driving a generator so that electricity can be produced. Vertical axis designs tend to have a lower profile on the seafloor and can be maintained more easily than horizontal axis designs since the device can be positioned closer to the surface. While they are less efficient than horizontal turbine designs in strong flows, they offer lots of flexibility in areas where the tidal current shifts.

Each of these ocean energy technologies operates on different mechanical principles, as discussed above, and so they perform differently depending on the conditions of the environment in which it is deployed. We hypothesized that among the wave and tidal energy technologies examined the performance and efficiency would vary based on the specific design and the locational environmental conditions in which they were built to operate and perform best. For example, we hypothesized that point absorber buoys will perform best in open ocean environments with constant wave motion, and that Oscillating Wave Surge Converters will excel in high-energy nearshore zones. Oscillating Water Column devices will be the most effective when integrated into existing structures and where wave action is predictable. For tidal technologies, we hypothesized that horizontal-axis turbines will perform best in strong directional tide flows and will produce larger energy quantities, while vertical-axis turbines will perform well in locations where the current will shift.

Materials and Methods

Wave and Tidal energy output was simulated and modeled using the System Advisor Model (SAM), which is a performance and financial simulation tool that was developed by the National Renewable Energy Laboratory (NREL). All simulations were run under the Marine Energy Model in the System Advisor Model, and the financial model was not applied.

For wave energy simulations, there were three types of converter technologies evaluated: a Wave Buoy, an Oscillating Surge Wave Converter, and a Floating Oscillating Water Column. Each converter was simulated across three resource sites, Humboldt Bay (California, United States), Jennette's Pier (North Carolina, United States), and PacWave North (Oregon, United States) which in total resulted in nine wave energy cases. For each one of these cases the correct wave resource file and converter were selected within SAM using default device and system parameters and the model was run to generate annual performance outputs. Summary statistics were recorded from the Metrics table, and annual energy production as a function of wave height (Hs) and energy period (Te) was extracted.

For tidal energy simulations, there were two types of converter technologies evaluated: a Vertical Axis Turbine and a Horizontal Axis Turbine. Each converter was simulated across three resource sites, Memorial Bridge (New Hampshire, United States), West Foreland (Alaska, United

States), and The Narrows (New York/New Jersey, United States), which in total resulted in six total tidal energy cases. For each one of these cases, the tidal resource input was configured with the Time Series option for each respective site, and default system parameters were selected within SAM. Summary statistics were recorded from the Metrics table, and annual energy production as a function of current speed (m/s) was extracted.

For both wave and tidal Converter technologies, summary data was extracted after running the SAM, such as capacity factor (%) and average power per device (kWh) for each technology and location.

Results

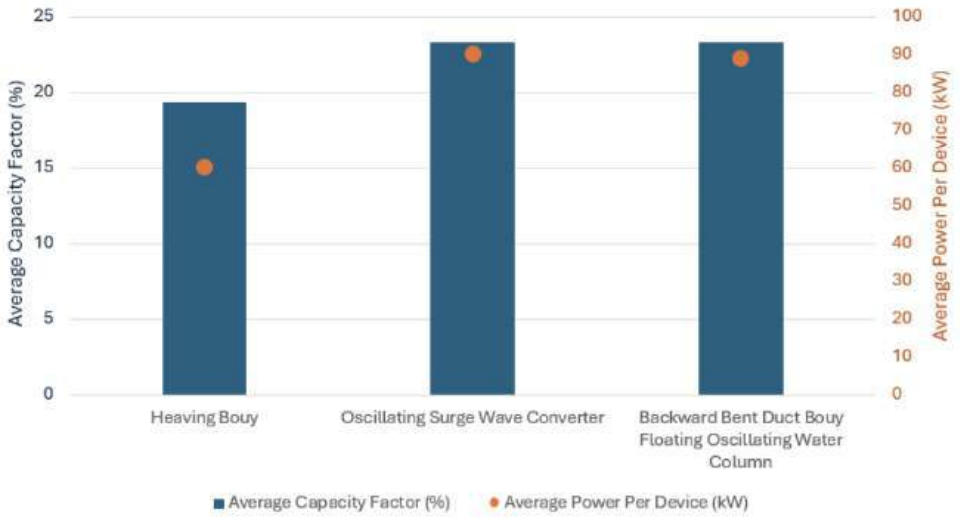


Figure 1. Average capacity factor and average power output comparison across wave energy converter technologies. Bar chart with overlaid circle plot showing average capacity factor (%) and average power output (kW) for three wave energy converter technologies: Wave Buoy, Oscillating Surge Wave Converter, and Floating Oscillating Water Column. Simulations were conducted using the System Advisor Model (SAM) with default device and system parameters.

To assess the efficiency vs the power output of wave energy converter technologies, the average capacity factor and average power per device were compared across the three technologies simulated. The Wave Buoy demonstrated the lowest performance of the three wave converters with an average capacity factor of 19% and an average power output of 60 kW (Figure 1). This indicates that while the device is able to capture energy within wave conditions, it utilizes a relatively small fraction of its potential power capture and produces limited power outputs. The Oscillating Surge Wave Converter and Oscillating Water Column performed similarly to one another, with both having average capacity factors of 23%, but the Oscillating Surge Wave Converter produced marginally higher average power at 90kW compared to 89 kW for the Oscillating Water Column (Figure 1). Both of these technologies, being nearly identical in capacity factors, suggest that they have comparable efficiency, and the slight power advantage of

the OSC shows that it is slightly greater at energy extraction. To evaluate how annual energy production varied across wave conditions and sites, annual energy output as a function of significant wave height (Hs) and energy period (Te) was examined for each technology type across Humboldt Bay (California, United States), Jennette's Pier (North Carolina, United States), and PacWave North (Oregon, United States).

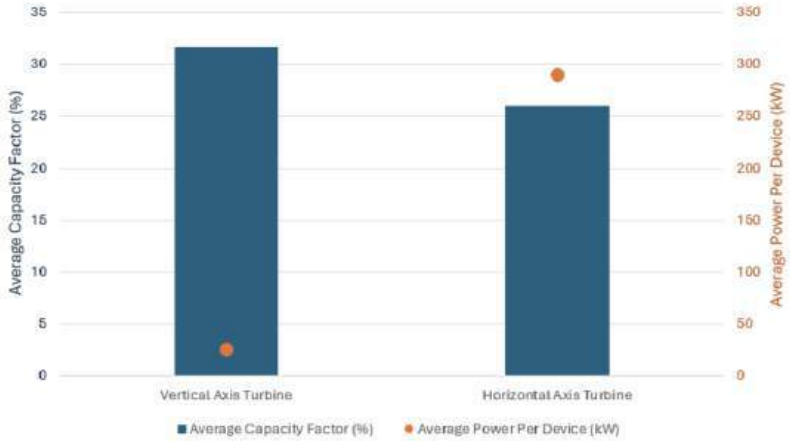


Figure 2. Average capacity factor and average power output comparison across tidal energy converter technologies. Bar chart with overlaid circle plot showing average capacity factor (%) and average power output (kW) for two tidal energy converter technologies: Vertical Axis Turbine and Horizontal Axis Turbine. Simulations were conducted using the System Advisor Model (SAM) with Time Series tidal resource inputs and default system parameters.

To compare the efficiency and power output of tidal energy converter technologies, the average capacity factor and average power per device were compared across the two technologies. The Vertical Axis Turbine had a higher average capacity factor of 32% compared to 26% for the Horizontal Axis Turbine; however, the Horizontal Axis Turbine produced substantially greater average power at 290 kW versus 26 kW for the Vertical Axis Turbine (Figure 2). This shows that even with a lower capacity factor for the Horizontal Turbine, meaning it is less efficient than the Vertical Turbine, the Horizontal Turbine extracts a far greater absolute energy output per device. To further analyze how annual energy production varied across tidal conditions, annual energy production as a function of tidal current speed was examined for both technologies across Memorial Bridge (New Hampshire, United States), West Foreland (Alaska, United States), and The Narrows (New York/New Jersey, United States).

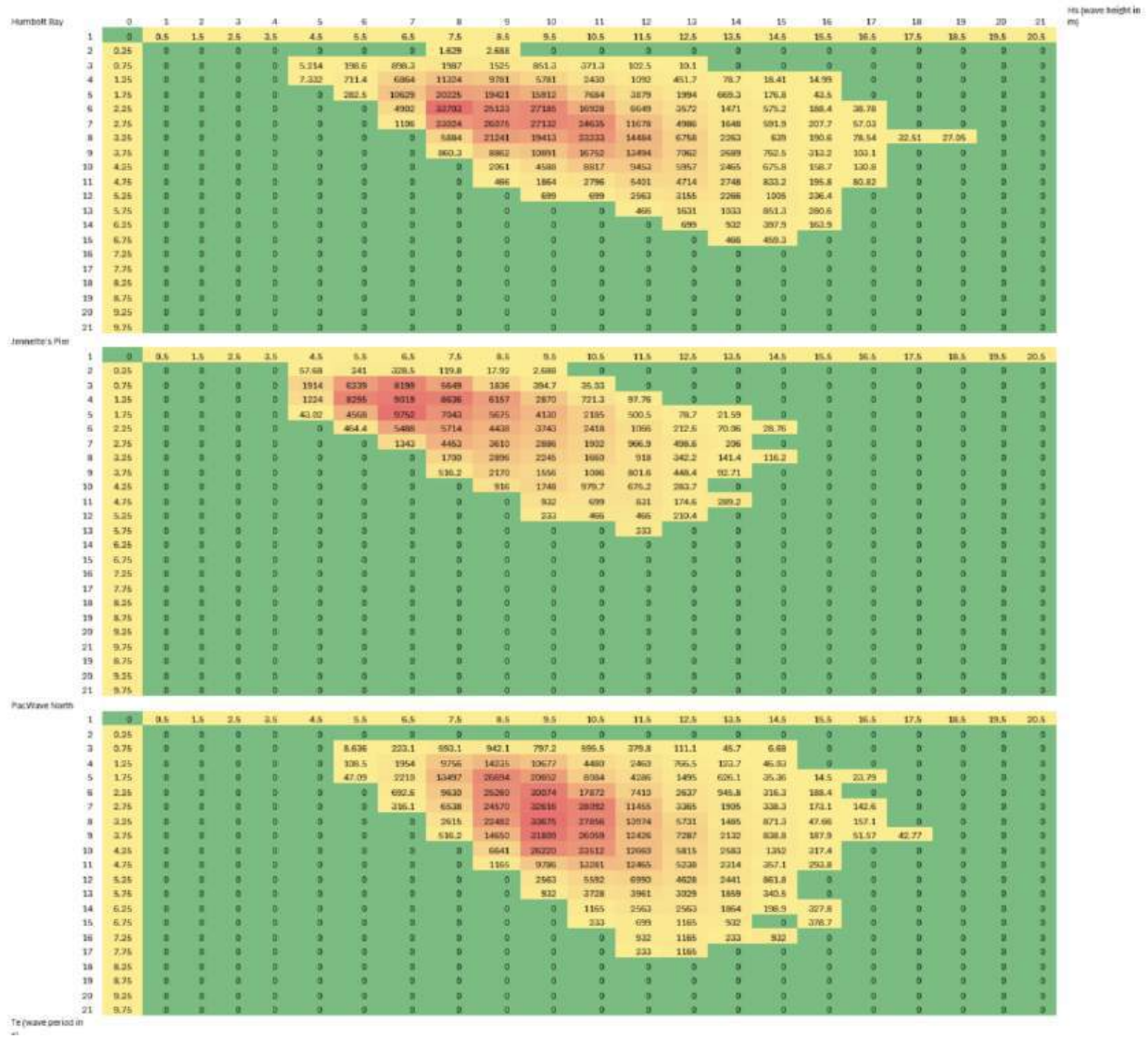


Figure 3. Annual wave energy production of the Wave Buoy converter across three resource sites. Heat maps showing annual energy production (kWh) as a function of significant wave height (Hs) and energy period (Te) for Humboldt Bay (California, United States), Jennette's Pier (North Carolina, United States), and PacWave North (Oregon, United States). Simulations were conducted using the System Advisor Model (SAM) with default device and system parameters.

For the Wave Buoy, PacWave North and Humboldt Bay were the strongest performing sites, with a peak annual energy output of 33,675 kWh at Hs = 10 m and Te = 8s, and 33,703 kWh at Hs = 8m and Te = 6s, respectively (Figure 3). While Jennette's Pier produced a much lower peak of 9,752 kWh at Hs = 7m and Te = 5s (Figure 3). This showcased that PacWave North and Humboldt Bay were more suitable locations for the Wave Buoy technology, whereas Jennette's Pier was not as well suited. High energy production across all three sites was tightly clustered within a narrow window of Hs = 6-12m and Te = 4-9s, and there was zero energy

production at wave heights below 5m or above 19m and at periods shorter than 2s or longer than 15s (Figure 3). At Humboldt Bay and Jennette's Pier, the energy distribution was skewed towards the lower wave periods and moderate heights, and PacWave North showed a more centrally distributed pattern, with moderate heights and periods (Figure 3).

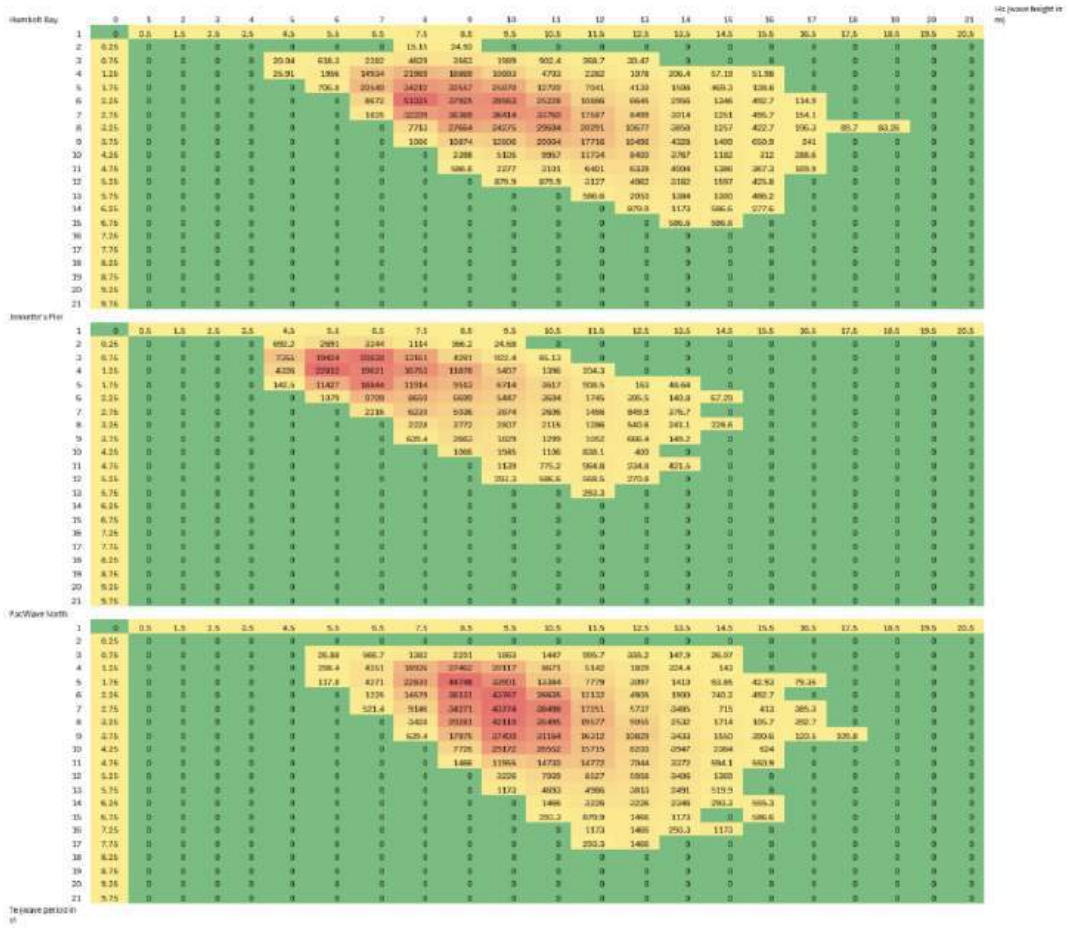


Figure 4. Annual wave energy production of the Oscillating Surge Wave Converter across three resource sites. Heat maps showing annual energy production (kWh) as a function of significant wave height (Hs) and energy period (Te) for Humboldt Bay (California, United States), Jennette's Pier (North Carolina, United States), and PacWave North (Oregon, United States). Simulations were conducted using the System Advisor Model (SAM) with default device and system parameters.

Next, the Oscillating Surge Converter was found to produce higher peak outputs than the Wave buoy across all sites. At Humboldt Bay, the OSC yielded the highest peak of 51,025 kWh at Hs = 8m and Te = 6s, followed by PacWave North at 44,748 kWh at Hs = 9m and Te = 5s, and lastly Jennette's Pier at 22,812 kWh at Hs = 6m and Te = 4s (Figure 4). This again showcased that PacWave North and Humboldt Bay were well-suited locations for the Oscillating Surge Converter technology, whereas Jennette's Pier was not as well suited. The high energy production at all three sites was tightly clustered between Hs = 6-12m and Te = 3-10s (Figure 4).

This shows that the OSC captures energy in similar wave condition ranges as the Wave buoy, but it does it with greater energy capture.

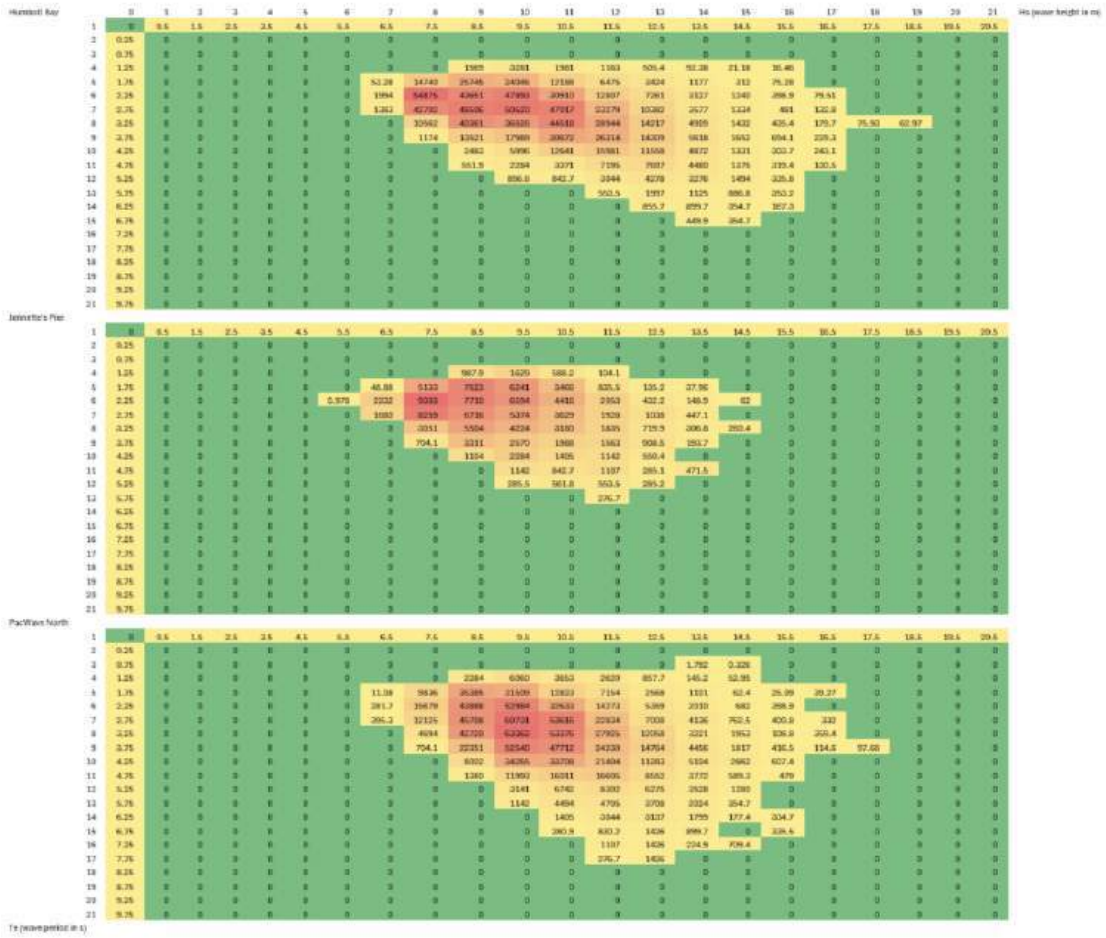


Figure 5. Annual wave energy production of the Floating Oscillating Water Column across three resource sites. Heat maps showing annual energy production (kWh) as a function of significant wave height (Hs) and energy period (Te) for Humboldt Bay (California, United States), Jennette's Pier (North Carolina, United States), and PacWave North (Oregon, United States). Simulations were conducted using the System Advisor Model (SAM) with default device and system parameters.

The Oscillating Water Column produced the highest peak energy outputs of all three wave technologies, with PacWave North recording the maximum of 63,362 kWh at Hs = 10m and Te = 8s, Humboldt Bay peaked at 54,875 kWh at Hs = 8m and Te = 7s, and Jennette's Pier only reached 9,303 kWh at Hs = 8m and Te = 6s (Figure 5). Unlike the Wave Buoy and OSC, which had similar high energy production zones, the Oscillating Water Column had high energy output from a slightly shifted range of Hs = 8-12m and Te = 5-10s, meaning it performed best in somewhat higher wave heights and longer periods (Figure 5). Energy distribution at Humboldt Bay and PacWave North was centrally clustered with a slight skew toward higher wave heights, while Jennette's Pier showed a more centrally distributed but low-output pattern (Figure 5).

Across all three wave energy technologies, Jennette's Pier consistently produced the lowest peak output regardless of the technology type. At Jennette's Pier, the Heaving Buoy peaked at 9,752 kWh, the Oscillating Surge Wave Converter at 22,812 kWh, and the Floating Oscillating Water Column at only 9,303 kWh (Figures 3,4,5). In contrast, Humboldt Bay and PacWave North produced significantly higher energy outputs of ~33,700 kWh for the Wave Buoy (Figure 3), 51,025 and 44,748 kWh for the OSC (Figure 4), and 54,875 and 63,362 kWh for the OWC (Figure 5).

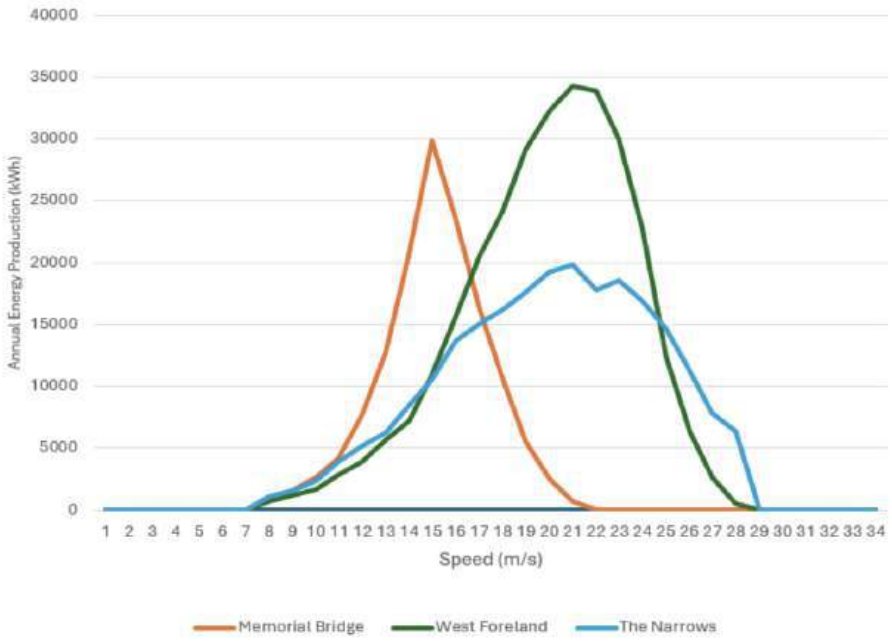


Figure 6. Annual tidal energy production of the Vertical Axis Turbine across three resource sites. Line graphs showing annual energy production (kWh) of the turbine array as a function of current speed (m/s) for Memorial Bridge (New Hampshire, United States), West Foreland (Alaska, United States), and The Narrows (New York/New Jersey, United States). Simulations were conducted using the System Advisor Model (SAM) with Time Series tidal resource inputs and default system parameters.

For the Vertical Axis Turbine, West Foreland was the strongest performing site with a peak energy output of 33,901 kWh at 22 m/s, followed by Memorial Bridge at 29,954 kWh at 15 m/s and The Narrows at 19,817 kWh at 21 m/s (Figure 6). Energy production began at 8 m/s at all sites, with zero energy output at speeds below this threshold (Figure 6). The high energy production varied by site, 14-16m/s at Memorial Bridge, 19-24 m/s at West Foreland, and 18-25 m/s at The Narrows (Figure 6). The energy curve at Memorial Bridge was skewed towards lower speeds with a sharp symmetric peak and a steep decline between 15-17 m/s, while West Foreland and The Narrows showed curves skewed towards higher speeds with broader plateaus before declining sharply above 24-27 m/s (Figure 6). All three sites had complete drop-offs to zero production above 29 m/s (Figure 6), showcasing that the Vertical Axis Turbine cannot perform at very high current speeds.

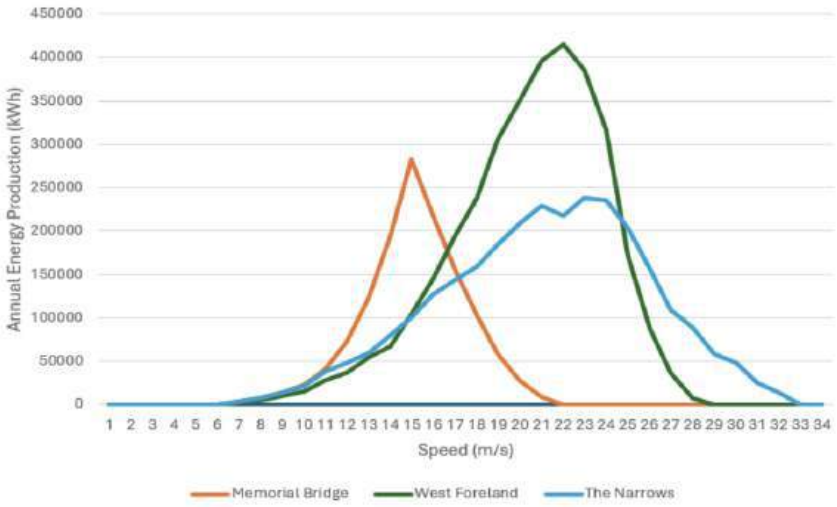


Figure 7. Annual tidal energy production of the Horizontal Axis Turbine across three resource sites. Line graphs showing annual energy production (kWh) of the turbine array as a function of current speed (m/s) for Memorial Bridge (New Hampshire, United States), West Foreland (Alaska, United States), and The Narrows (New York/New Jersey, United States). Simulations were conducted using the System Advisor Model (SAM) with Time Series tidal resource inputs and default system parameters.

The Horizontal Axis Turbine produced dramatically higher energy outputs than the Vertical Axis Turbine across all three sites, with West Foreland again being the strongest site at a peak of 415,392 kWh at 22 m/s, followed by Memorial Bridge at a peak of 282,618 kWh at 15 m/s and The Narrows with 238,245 at 23 m/s (Figure 7). These peak values are approximately 10-12 times greater than the output for the Vertical Axis turbine, which peaks at the same sites. Energy production began at 7 m/s across all sites (Figure 7). The high energy zones were similar to that of the Vertical Axis Turbine, 14–16 m/s at Memorial Bridge, 19–24 m/s at West Foreland, and 20–25 m/s at The Narrows (Figure 7). The Horizontal Axis Turbine maintained high output across a broader speed range before declining, particularly at West Foreland and The Narrows where the curve exhibited a more gradual plateau, rather than the sharp symmetric peak seen at Memorial Bridge (Figure 7).

Across both tidal technologies, West Foreland was the strongest site for both converter types, producing peaks of 415,392 kWh for the Horizontal Axis Turbine and 33,901 kWh for the Vertical Axis Turbine (Figures 6,7). Memorial Bridge ranked second in peak output, 282,618 kWh for the Horizontal Axis Turbine and 29,854 kWh for the Vertical Axis Turbine (Figures 6,7). The Narrows produced the lowest peaks of the three sites at 238,245 kWh and 19,817 kWh for the Horizontal and Vertical Axis Turbines, respectively (Figures 6,7). The performance gap between sites was proportionally consistent across both technologies, with West Foreland outperforming The Narrows by approximately 74% for the Horizontal Axis Turbine (415,392 vs. 238,245 kWh) and 71% for the Vertical Axis Turbine (33,901 vs. 19,817 kWh). The Horizontal Axis Turbine remained the superior technology by a factor of approximately 10–12 times in

absolute output across all sites, making it the more effective choice for high-energy tidal environments.

Discussion

Wave and Tidal energy systems represent two of the most promising, however, underused forms of renewable energy, even though offering predictable and consistent resource availability similar to that of solar and wind renewable technologies. This study evaluated the simulated annual energy performance of three wave energy converter technologies, the Wave Buoy, Oscillating Surge Converter, and Oscillating Water Column, across three coastal sites. In addition, simulated the annual energy performance of two tidal energy converter technologies, the Vertical Axis Turbine and Horizontal Axis Turbine, across three tidal sites. All of these simulations were done using the System Advisor Model, and the results demonstrated a clear difference in both technology performance and site suitability, with Pacific coastal sites outperforming the Atlantic site for wave energy, and Alaskan tidal resources outperforming northeastern United States sites for tidal energy.

Among the three wave energy technologies evaluated, the Oscillating Water Column had the highest peak energy values at the two strongest resource sites, reaching 63,362 kWh at PacWave North and 54,865 kWh at Humboldt Bay, showcasing that it was the most capable technology evaluated for high-energy Pacific wave environments. However still, the identical capacity factors of the Oscillating Surge Converter and Oscillating Water Column show that the two technologies are comparably efficient. The high energy output of the Oscillating Water Column at strong resource sites shows the suitability for the wave and climate characteristics of the Pacific coast, specifically the moderate to high wave heights and longer energy periods. The Oscillating Surge Converter, while producing slightly lower energy outputs, achieved its highest output at Humboldt Bay and PacWave North, showcasing that it is also suited for Pacific coast environments. This shows that both the OWC and OSC perform efficiently and are well-suited for Pacific coast wave environments, as they both performed their best in these locations.

The Wave Buoy consistently underperformed compared to the other two wave technologies, with the lowest average capacity factor at 19% and the lowest average power output at 60 kW. Therefore, the Wave Buoy achieved moderately lower peak outputs than the OSC at Humboldt Bay and PacWave North at the best Hs and Te conditions. In addition, the Wave Buoy captured energy in a narrower range of productive high-energy capture than the OSC, which makes it more sensitive to deviations from the optimal window of Hs = 6-12 m and Te = 4-9 s. This sensitivity to wave conditions is consistent with the physical operating principle of Wave Buoys, which rely on the resonance between the device and waves to maximize capture. However, when wave characteristics fall outside this range, the energy extraction drops very harshly. This implies that Wave Buoys may be better suited for deployment in consistent wave environments, where conditions regularly fall within the productive range, rather than sites with more variable conditions.

Across all of the wave technologies, the site Jennette's Pier produced substantially lower energy outputs than the two Pacific sites (PacWave North and Humboldt Bay), with peak outputs at 55% lower for any given technology. This performance gap could be explained by the differences in wave climates between the Atlantic and Pacific coasts of the United States. The Pacific coast is characterized by long swells generated by storm conditions, producing moderate to high wave heights and longer energy periods that align well with the technologies evaluated. Jennette's Pier, located on the Outer Banks of North Carolina, experiences a more variable wave climate driven by shorter wind waves, seasonal hurricane swell, and generally lower wave heights than that of the Pacific coast, which all collectively limit energy production regardless of technology type. This suggests that a large-scale wave energy deployment in the United States would be most viable along the Pacific coast, especially for the converter technologies simulated in this study.

For tidal energy, the Horizontal Axis Turbine had extremely superior absolute energy output compared to the Vertical Axis Turbine, producing approximately 10-12 times greater peak energy across all three sites. This is despite having a lower average capacity factor than the Vertical Axis Turbine, where the Horizontal Axis Turbine was 26%, and the Vertical Axis Turbine was 32%. This is explained by the substantially higher rated average power per device of the Horizontal Axis Turbine Design, which can extract far greater energy per unit of tidal current than the Vertical Axis Turbine, even when operating at a smaller fraction of the theoretical max. This suggests that for any large utility-scale tidal energy generation, the Horizontal Axis Turbine would be the preferred technology, while the Vertical Axis turbine may find use in lower current environments, or smaller-scale deployments, where the more efficient and consistent profile is more advantageous.

The site-based differences in tidal energy output were consistent across both technologies and closely tied to the current characteristics of each location. West Foreland, Alaska, was the strongest performing site, with peak outputs of 415,392 kWh and 33,901 kWh for the Horizontal and Vertical Axis Turbines, respectively. This could be attributed to the naturally fast and steady tidal currents that occur at the site, and the efficient production zones at West Foreland were 19-24 m/s for both technologies, which aligns well with the higher speed operational ranges of the turbines. Memorial Bridge, while ranking second in peak output, had considerably lower speeds for its maximums at a speed of 15 m/s for both technologies, which reflects slower tidal currents in the Piscataqua River in New Hampshire. This lower optimal speed at Memorial Bridge suggests that it could still be a viable tidal resource site, but it produces substantially less energy. The Narrows ranked lowest in peak output at 238,245 kWh and 19,817 kWh for the Horizontal and Vertical Axis Turbines, respectively, despite having a moderately strong tidal resource and one of the broader efficient speed zones across sites at 20-25 m/s. This could be explained by the fact that The Narrows reaches high current speeds; the distribution of these speeds does not hit a consistent range throughout the whole year, as well as West Foreland. Site resource quality and current speed are the primary factors of differences between energy outputs rather than any technology-specific factors, as seen in wave energy. This is shown through the

proportionally consistent performance gap between West Foreland and The Narrows across both technologies, approximately 74% for the Horizontal Axis Turbine and 71% for the Vertical Axis Turbine. Therefore, for tidal energy, site selection is a larger factor in determining viability for energy outputs, whereas for Wave technologies, matching technology type to the site selection was critical.

Several technology-specific improvements could be made in future iterations to address the specific performance limitations previously identified in this study. For the Wave Buoy, the complete absence of energy production at wave heights below 5 m and periods below 2 s suggests that involving low wave sensitivity mechanisms (lighter floats or enhanced hydraulic damping systems) could extend the operational range. By doing so, the device could capture energy from smaller waves, which would boost energy outputs, especially at lower resource sites like Jennette's Pier. Similarly, introducing variable geometry floats or adjustable damping systems could widen the high-energy wave window beyond the narrow spot of Hs = 6-12 m and Te = 4-9 s, boosting energy production. For the OSC, upgrading hydraulic transmission efficiency and optimizing paddle geometry for lower wave heights and shorter periods could improve performance at sites like Jennette's Pier where the wave climate does not consistently reach the optimal range for this technology. For the OWC, it could similarly benefit from adjustable air chamber valves or multiple entry ducts to extend its effective energy capture to slightly lower wave heights of 6- 7 m and periods of 4-5 s, which would improve performance at sites where the climate does not reach the higher optimal range. More broadly, a hybrid wave energy deployment, combining multiple converter types, could provide a more extensive coverage of operational windows and improve total energy capture at sites where no single technology performs optimally, i.e., in sites with inconsistent wave heights or periods.

For Tidal energy, the Horizontal Axis Turbine's sharp energy decline at the highest tidal speeds (occurring between 15-17 m/s at Memorial Bridge, 24-26 m/s at West Foreland, and 25-27 m/s at The Narrows) suggests that reinforcing rotor blade structures and implementing active pitch control systems could sustain energy production into the higher speed range and reduce losses at sites with strong tidal currents. At the lower end of the speed range, implementing variable pitch blades or starter fin mechanisms could improve the energy capture below the 7 m/s threshold. Both of these upgrades would improve the energy capture output by widening the range for efficient energy capture. For the Vertical Axis Turbine, given the extremely lower peak outputs than the Horizontal Axis Turbine across all three sites, it is recommended that the Vertical Axis Turbine should not be used for large-scale tidal energy deployment. Rather, future tidal energy development efforts would be better directed towards optimizing and scaling Horizontal Axis Turbine technology, rather than investments in Vertical Axis Turbine deployments.

Several limitations of this study should be considered when interpreting these results. All simulations were conducted using the default device and system parameters within SAM, meaning that there was no device-specific tuning, no array configurations, and no installation constraints. Real-world deployments would involve optimization of device parameters to local

wave or tidal conditions, which could alter the performance outputs and rankings observed here. In addition, the wave energy matrices and tidal speed curves represent the annualized outputs from historical resource data and do not capture the interannual variability in wave or tidal conditions that could influence the long-term energy production. Additionally, these simulations were run with the financial modeling turned off in SAM, and transmission infrastructure and grid integration constraints were not evaluated, which are all critical factors in real-world deployment decisions. This could disproportionately affect the feasibility of the more remote sites like West Foreland, Alaska, despite being the strongest tidal resource. Future studies should examine these technologies under optimized device parameters per site, consider using array deployments rather than individual devices, and incorporate logistical and economic modeling to evaluate the true cost of energy at each site. In addition, future studies should also expand the number of locations evaluated, particularly an analysis of additional Atlantic locations, and introducing Gulf Coast locations would provide a better picture of US marine energy potential.

In summary, this study demonstrated that among wave energy technologies, the OWC produced the greatest peak energy output at high-resource-available sites, like those in the Pacific coast, and the OSC also achieved comparable efficiency and strong outputs in Pacific sites. The Wave buoy underperformed both alternative wave converter technologies across all metrics in all locations. For tidal energy, the Horizontal Axis Turbine far exceeded the Vertical Axis turbine in peak energy outputs at all three sites, with West Foreland emerging as the most productive tidal resource location evaluated. Across both marine energy categories, the Pacific wave sites and the Alaskan tidal site performed the strongest compared to their Atlantic coast counterparts. There is also a fundamental difference in the approach of how wave and tidal energy systems should be deployed. For wave energy, both site selection and technology choice are equally important, making sure the converter types play to the advantages of that site. For tidal energy, site resource quality was the dominant factor in determining energy output. All in all, these findings highlight the promising future for marine energy deployment, and through specific site-based improvements and deployment, where technology selection should be carefully matched to local resource characteristics, marine energy could develop into a viable renewable energy resource in the United States.

Works Cited

Burton, Tony, et al. Wind Energy Handbook. 2nd ed., Wiley, 2011.

Drew, Ben, et al. "A Review of Wave Energy Converter Technology." Proceedings of the Institution of Mechanical Engineers, Part A, vol. 223, no. 8, 2009, pp. 887–902. https://doi.org/10.1243/09576509JPE782.

Ellabban, Omar, et al. "Renewable Energy Resources: Current Status, Future Prospects and Their Enabling Technology." Renewable and Sustainable Energy Reviews, vol. 39, 2014, pp. 748–764. https://doi.org/10.1016/j.rser.2014.07.113.

Falcão, António F. de O. "Wave Energy Utilization: A Review of the Technologies." Renewable and Sustainable Energy Reviews, vol. 14, no. 3, 2010, pp. 899–918. https://doi.org/10.1016/j.rser.2009.11.003.

Global Wind Energy Council. Global Wind Report. 2023, https://gwec.net/global-wind-report-2023/.

Intergovernmental Panel on Climate Change. Climate Change 2021: The Physical Science Basis. IPCC, 2021, https://www.ipcc.ch/report/ar6/wg1/.

International Energy Agency. Solar Energy: Tracking Report. https://www.iea.org/reports/solar-pv.

Kabir, Ehsanul, et al. "Solar Energy: Potential and Future Prospects." Renewable and Sustainable Energy Reviews, vol. 82, 2018, pp. 894–900. https://doi.org/10.1016/j.rser.2017.09.094.

Khan, M. J., et al. "Ocean Tidal Energy: Review of Technologies and Environmental Impact." Renewable and Sustainable Energy Reviews, vol. 12, no. 9, 2008, pp. 2386–2403. https://doi.org/10.1016/j.rser.2007.06.011.

Paish, Oliver. "Small Hydro Power: Technology and Current Status." Renewable and Sustainable Energy Reviews, vol. 6, no. 6, 2002, pp. 537–556. https://doi.org/10.1016/S1364-0321(02)00006-0.

Torre-Enciso, Yolanda, et al. "Wave Energy in Europe: Current Status and Perspectives." Renewable and Sustainable Energy Reviews, vol. 13, no. 8, 2009, pp. 2126–2131. https://doi.org/10.1016/j.rser.2009.01.012.

Did the Expansion of Telemedicine in India (2020–2023) Correlate With a Reduction in Non-urgent Outpatient Department (OPD) Visits at Public Hospitals?

By Sarvika Chaudhary

Abstract

The rapid emergence of telemedicine in India since 2020 has transformed access to healthcare significantly, particularly in the field of outpatient services that are not urgent. The current paper concentrates on the correlation between the growth of telemedicine (20202023) and the reduction of the non-urgent visits to OPD in state hospitals, assessing the efficiency of digital health interventions in overcrowding reduction in outpatient hospitals.

The study is based on aggregated data about OPD utilisation of Health Management Information System (HMIS) and data regarding teleconsultation of official reports of eSanjeevani. Time-series and correlation-based analysis were used to analyze the differences between pre- and post-implementation periods.

The findings show that telemedicine consultation numbers rose to over 100 million in 2023, as compared to about 1 million in 2020. Meanwhile, even the visits to the public OPDs were reduced to 980 million in 2020 and recuperated to about 1.05 billion in 2023, which is not yet representative of population growth and the increased healthcare demand. States that experienced high OPD burdens, including Tamil Nadu, Kerala, and Rajasthan, demonstrated a strong presence of telemedicine and a sustained OPD pressure. Most of the teleconsultations discussed were follow-up appointments, chronic illnesses, minor acute conditions, and mental health support.

These results are congruent with the hypothesis that telemedicine has partially replaced non-urgent physical OPD care, alleviating congestion with no discontinuity, but lacks causality due to confounding by the pandemic and aggregation of data.

Keywords telemedicine; eSanjeevani; outpatient department (OPD); public hospitals; healthcare utilisation; India.

Introduction

The COVID-19 crisis has put an unprecedented strain on the national healthcare systems of all countries, especially in India where millions of people still visit overcrowded public OPDs even under the conditions of limited mobility. The government responded by quickly expanding eSanjeevani, a nationwide telemedicine platform that operates on a hub-and-spoke telemedicine platform, particularly to redirect non-urgent outpatient care out of physical hospitals and instead to remote consultations. This brought up an important empirical issue, namely, whether increased access to telemedicine leads to a decrease in physical OPD utilisation or simply an increase in an additional stratum to the current healthcare demand? Although the intended scale of telemedicine implementation in India is significant since 2020, there is a lack of empirical studies that directly analyze this relationship. The gap discussed in this paper is the analysis of the correlation

between the increase of telemedicine and the increase/decrease of visits to non-urgent OPD in the period between 2020 and 2023 based on the use of aggregated HMIS and eSanjeevani data to assess whether telemedicine has been an alternative or a complement to physical outpatient care.

Rapid development of telemedicine has been extensively studied with references to healthcare access, continuity and efficiency of care provision, especially after the COVID-19 pandemic. The previous literature emphasizes the importance of telemedicine to enhance access to underserved groups, minimize traveling costs and opportunity costs, and aiding routine care in times of health system disruption (Whitelaw et al., 2022; Singh and Kapoor, 2025). Digital health platforms have been increasingly placed in the Indian context as a scalable solution to overcome constraints in the structure of the public healthcare system (World Scientific News, 2021).

A number of studies have examined the national telemedicine platform eSanjeevani in particular with respect to design, implementation and patterns of utilisation. Published in the Oxford Open Digital Health, research has documented the rapid scale-up of eSanjeevani and its integration in the public delivery of healthcare, its hub-and-spoke nature and focus on outpatient consultations (Basu et al., 2025). Likewise, JMIR Medical Informatics research studies the adoption of telemedicine by providers and its operational viability, with more clinicians and patients showing greater acceptance of telemedicine during and after the pandemic (Whitelaw et al., 2022).

Other sources have also discussed the user experiences, acceptability and system level preparedness to adopt telemedicine. The literature in JMIR Formative Research and INHS shows overall positive attitudes towards telemedicine in the context of follow-up and minor health issues, but also finds the lack of digital literacy and infrastructure a barrier to telemedicine implementation (Sharma et al., 2025; Singh & Kapoor, 2025). Research in the Indian Journal of Pharmacy Practice further indicates that telemedicine can be used as a supportive tool in the management of medication counselling and routine outpatient contacts in state-run healthcare institutions (IJPP, 2021).

Regardless of this accumulating body of literature, there are very few studies that directly analyse the association between the growth of telemedicine and the use of physical outpatient departments (OPD). Current studies on OPD-specific issues in India have largely focused on the quality of services, patient satisfaction, and waiting times without a significant number of studies evaluating the system by comparing visit numbers before and after the introduction of telemedicine (World Scientific News, 2021).

This study fills an important gap in the literature by specifically examining the utilisation of non-urgent OPD. Knowing whether telemedicine is associated with lower use of physical OPD services is significant in appraising the efficiency of the health system and to guide policy actions to promote increased use of telemedicine. Evidence of the relationship between the expansion of telemedicine and alterations in the utilisation of OPDs can underpin plans that would encourage telemedicine as a feasible substitute to routine consultation in the Indian public healthcare system (Basu et al., 2025; Whitelaw et al., 2022).

Research Gap and Originality

Although previous studies have considered the rate of adoption of telemedicine, patient satisfaction, and qualitative evaluation of digital health services, not many studies have been able to directly evaluate whether telemedicine, in fact, replaces physical OPD visits in resource-constrained countries like India. The majority of literature available on OPD-specific problems is dedicated to assessing the quality of services, patient satisfaction, or waiting time, without assessing whether the number of visits to clinics increases or stays the same after the growth of telemedicine. It is especially large in low- and middle-income countries where the public healthcare sector experiences persistent capacity gaps and digital interventions into health are actively encouraged as efficiency strategies. This paper addresses this gap by synthesising official government telemedicine and OPD utilisation data to test directly the substitution hypothesis at the national and state levels in India.

Methods

Alternative Approaches Considered

Three principal methodological designs were considered prior to the choice of a final design. A case study of a single, tertiary centre at a hospital level was the first option to be considered that would have given granular visit-level data but would not have given national generalisability and would have needed institutional access, which is unavailable at this level of study. The second methodology we have taken was the state-level regression model with covariates of the hospitals including bed capacity, specialists density, and internet penetration. Though statistically more rigorous, this method would have necessitated microdata and research collaborations not feasible within the scope of a school-based research project. The third methodology, national time-series and correlation analysis of government data publicly available, was chosen due to the combination of most favourable data availability, national generalisability, and analytical tractability to the scale of the project. Although such a methodology cannot be used to determine causality, it is highly appropriate in determining co-occurring trends and testable hypotheses.

Study Design and Analysis Plan

This research adopted a quantitative and observational design, with the analysis carried out across six structured steps. The first involved obtaining OPD data from the HMIS (Ministry of Health and Family Welfare) and eSanjeevani usage reports covering 2019 to 2023, as these represent the only publicly available official sources offering systematic annual reporting at both national and state levels. The second step required standardising annual indicators, specifically total public OPD visits and telemedicine consultations measured in millions, to enable longitudinal comparison, since standardisation ensures that absolute and relative trends remain comparable across years. Third, trends were visualised through line graphs and bar charts to determine whether telemedicine growth and OPD stagnation co-occur temporally, with

visualisation preceding statistical analysis to confirm whether an observed pattern merits further quantitative investigation. Fourth, a state-level disaggregation was conducted to examine whether states with greater telemedicine adoption also exhibited sustained or reduced OPD pressure relative to demand expectations, thereby testing whether the national trend holds geographically. Fifth, the nature of care delivered through telemedicine was profiled by reviewing official eSanjeevani records and peer-reviewed literature on teleconsultation categories, as identifying the overlap between telemedicine cases and non-urgent OPD types is essential to assessing the plausibility of substitution. Finally, the results were interpreted in light of causality constraints, pandemic confounders, and data aggregation limitations, with policy implications discussed, since honest acknowledgment of inferential boundaries is integral to producing credible and valid research findings.

Resources and Support

This project utilized publicly available data: Health Management Information System (HMIS) data through the national health portal of the ministry of health and family welfare, and eSanjeevani usage data through official government releases and ministry press releases. Google Scholar and PubMed were identified as the sources of peer-reviewed literature through specific search queries that included terms like: telemedicine India, eSanjeevani, outpatient department utilisation and digital health services. Only full-text, English-language accessible studies with a publication date of 2019-2025 were included. The standard spreadsheet tools were used to organise data, build tables and visualise trends. This was an independent project that was not supported by any outside mentor or institution.

Study Design

The current study has had a quantitative, observational design to investigate the relationship between the growth of telemedicine services and the shift in the utilisation of non-urgent outpatient departments (OPD) in India in 2020-2023. The research was aimed at investigating the dynamics in the number of national-level physical outpatient visits and telemedicine use in the public healthcare system. It was not an experimental study but instead sought to determine associations through time and not causality.

Sources of Data and the choice of literature

The official government websites reporting on healthcare utilisation in India were used to get the data. The data of the outpatient department visits was obtained through the Health Management Information System (HMIS) of the Ministry of Health and Family Welfare, which reports the current OPD attendance per year at all the public health care institutions. Official records and those published in the form of statistics were used to obtain data on the utilisation of telemedicine in reference to eSanjeevani, the national telemedicine platform of the Government of India. Peer-reviewed literature (relevant journals such as Oxford Open Digital Health, JMIR Medical Informatics, JMIR Formative Research, INHS, Indian Journal of Pharmacy Practice,

and World Scientific News) were **consulted to support the interpretation of the observed trends.**

Data Analysis

Descriptive analysis was used to summarise annual trends in OPD visits and telemedicine consultations. A time-based comparison examined changes in OPD utilisation before and after the expansion of telemedicine services. Correlation analysis was used to explore the relationship between telemedicine usage and non-urgent OPD attendance over time.

Ethical Considerations

All data used in this study were publicly available and reported in aggregate form. No individual-level or identifiable information was accessed, and therefore ethical approval was not required.

Results and Analysis: Telemedicine Expansion and Outpatient Department (OPD) Utilisation in India

1. The significance of OPD Services in the Indian Public Health System.

The backbone of the healthcare delivery in the Indian public health system is outpatient department (OPD) services. In comparison to inpatient services, which provide care to a smaller number of cases that are severe, OPDs bear the majority of healthcare demands, such as regular consultations, follow-ups, chronic illnesses, maternal and child visits, and reporting of early symptoms. Before the COVID-19 pandemic, India reported an approximate 1.2 billion public OPD visits per year, highlighting the fact that the system was based on physical, hospital-based outpatient services.

The prevalence of OPD services is also seen in the surveillance data gathered by the Ministry of Health and Family Welfare (MoHFW). In April to June 2020 alone, India had over 113 million OPD visits, and about 6.7 million IPD admissions meaning that over 94% of reported healthcare contacts were at the outpatient level. This dependence on physical access to OPD services implied that a major threat to continuity of care would occur with non-urgent and planned medical demands.

2. Trends in the use of OPD in the country before and after the expansion of telemedicine.

2.1 Pre-pandemic OPD Patterns

The growth in the use of OPD in India up to 2020 was linked to the growing population, transition to chronic diseases, and the rising level of awareness of healthcare services. Public hospitals were commonly over-crowded with long waiting time, and short consultation time especially in high population states. In spite of these obstacles, OPDs were the default point of care of most citizens as they were affordable and available.

2.2 OPD Decrease in the COVID-19 Pandemic.

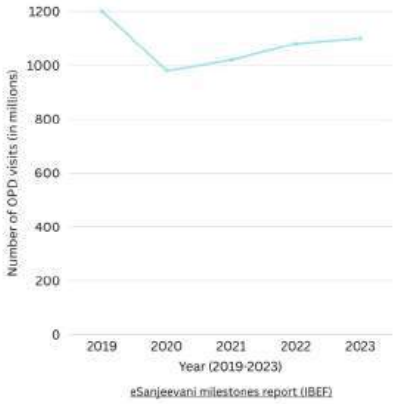
During 2020, the attendance at OPDs was decreasing drastically due to lockdowns, restrictions on mobility, and fear of being infected. According to national estimates, this would be a decrease in the number of visits to public OPDs, which amounted to 1.2 billion in 2019 and 980 million in 2020, or about 18% decrease. This decline was however not an indication of reduced healthcare needs but it reflected deferred or diverted care.

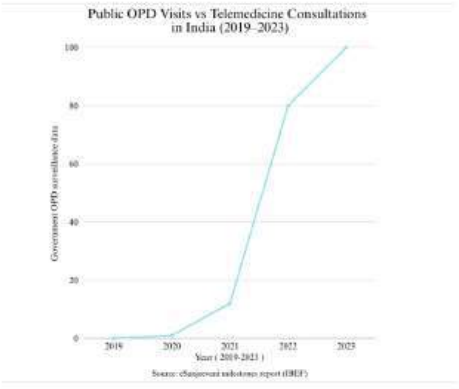
2.3 Telemedicine as Substitute to OPD.

At the same time there was rapid growth of telemedicine services. The eSanjeevani platform, led by the government, which was introduced nationwide in late 2019 was aggressively scaled during the pandemic. The level of telemedicine consultations grew insignificant before 2020 and reached about 1 million consultations in 2020, then started growing exponentially over the following years.

As of 2021, telemedicine consultations had reached some 12 million, 80 million in 2022, and 100 million consultations in 2023, as per MoHFW releases. Significantly, most of these consultations consisted of standard OPD level care, such as follow-up, management of chronic diseases, and minor symptom evaluation.

Graph1. National Trends in Public OPD Visits and Telemedicine Consultations in India (2019–2023)





Graph 1 (Line Graph): Public OPD visit and telemedicine (consultation) per year, with a significant increase in telemedicine and slower recovery in OPD after 2020.

3. Burden of OPD and Telemedicine Relevance on a State Level.

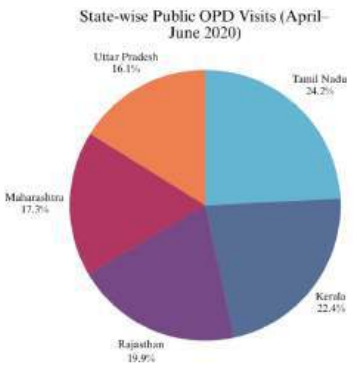
3.1 Disparities in OPD Loading among the States.

Surveillance data at the state level reveals that there are wide differences in the use of OPD in India. In the months of April-June 2020, numerous states have recorded unprecedentedly high volumes of OPDs in the face of pandemic limitations. These included:

14.7 million OPD visits in Tamil Nadu.
Kerala: 13.6 million visits to OPD.
Rajasthan: 12.1 million OPD visits
Maharashtra: 10.5 million visits in OPD.
Uttar Pradesh: 9.8 million OPD visits

These numbers indicate that outpatient care demand continued even at crisis levels especially in those states that had a better public health infrastructure and higher reporting coverage.

Graph 2. State-wise Public OPD Visits (April–June 2020)



3.2 Pressure on Disease Surveillance and OPD

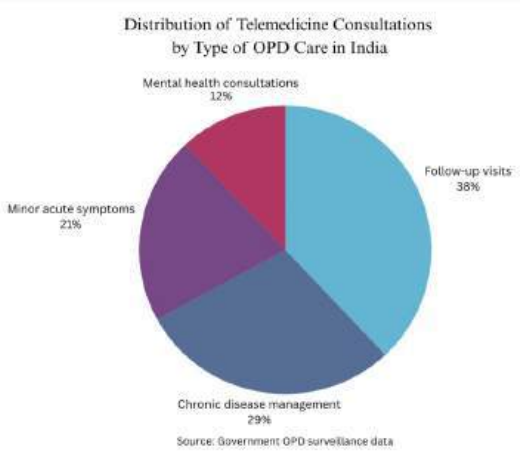
The same state-wise data show a high reporting of the influenza-like illness (ILI), severe acute respiratory infection (SARI), and a case of prolonged fever, all of which are generally handled at the OPD level. Indicatively, Tamil Nadu has recorded an excess of 1 million SARI cases by June 2020, and Maharashtra and Kerala have recorded a high ILI burden. Such circumstances necessitated frequent visits and follow-ups, which put a long-term strain on OPD services.

The telemedicine implementation in these high-burden states enabled the health systems to handle non-severe cases at a distance without physical congestion, and continuity of care.

4. Character of OPD Care Changed to Telemedicine.

- The adoption of Telemedicine in India has been very selective as opposed to global. Peer-reviewed studies and official reports all demonstrate that telemedicine services are used to replace non-urgent and planned OPD care, but not emergency and inpatient care.
- According to the analysis of the records of teleconsultation, the largest portion of telemedicine use has the following categories:
- Follow-up appointments, such as post-treatment check-ups and drug re-evaluations.
- Chronic illnesses, including diabetes, high blood pressure and thyroid illnesses.
- Mild acute symptoms, such as respiratory, gastrointestinal, and skin problems.
- Mental health consultation especially anxiety and depression.

Table 3. Distribution of Telemedicine Consultations by Type of OPD Care



This distribution reinforces the argument that telemedicine functions as an alternative OPD delivery channel, absorbing routine consultations that would otherwise require physical hospital visits.

5. correlation between Telemedicine Expansion and OPD Utilisation.

When both national and state-level trends are considered, a common trend can be observed. Although the number of OPD visits dropped significantly in 2020, it started increasing slowly in the following years, even though regular healthcare services were already provided again. Conversely, telemedicine consultations were still gaining exponentially although mobility had been eased.

This deviation indicates that telemedicine was not a short-term crisis solution but helped to structure the change in outpatient care utilisation. Some of the usual OPD demand seems to have been permanently diverted to telemedicine platform, specifically planned and non-urgent care.

Though the causality can not be proved definitively based on aggregated data, the fact that the growth of telemedicine and the slower pace of recovery of OPD occur in a temporal order

provides support to the hypothesis that telemedicine is replacing physical OPD visits and not merely supplementing it.

6. Value of an OPD-Centered Telemedicine Model.

It is especially important to concentrate on the utilisation of OPD to determine the actual impact of telemedicine in India. Congestion in OPDs has been long related to long waiting times, poor quality of consultations, burnout of providers, and higher risk of infectious disease transmission. Telemedicine assists in relieving such systemic strains by reallocating non-urgent consultations out of physical hospitals.

Moreover, telemedicine with OPD could promote more timely care-seeking behavior by reducing opportunity costs (travel time, lost wages, and discomfort (overcrowding) associated with going to the doctor). This is especially important to the elderly, women, and those with chronic illnesses needing regular follow-ups.

Policy-wise, facts that telemedicine can decrease physical OPD burden reinforce the argument that the digital health infrastructure should be maintained. It also promotes the use of telemedicine as part of the regular primary care provision and not as an emergency intervention.

Limitations, Lessons Learnt, and Future Work

This project reinforced several important lessons. Working with government health databases such as HMIS and eSanjeevani requires careful attention to reporting conventions and year-on-year consistency, and time-series analysis, while valuable for identifying systemic change, cannot replace experimental designs when causal claims are needed. Correlational findings must always be interpreted with explicit acknowledgment of confounders, and defining constructs like “non-urgent OPD care” precisely is essential since proxy definitions introduce ambiguity that must be transparently acknowledged.

If this research were extended, future work would prioritise collecting state-level panel data for fixed-effects regression, surveying patients directly to gather individual-level evidence for the substitution hypothesis, and accessing hospital-level microdata to enable stronger causal estimation. Extending the time series beyond 2023 and incorporating private telemedicine data would further strengthen the analysis.

Limitations

There are some limitations to this study and the findings can be interpreted with these limitations in mind. To begin with, the analysis is based on aggregated data at the national and state levels, which does not allow following the behaviour of a particular patient. Consequently, the research is not able to establish whether a particular telemedicine consultation directly substituted a physical OPD visit or whether more fundamental changes in healthcare-seeking behaviour were affecting the trends.

Second, the research is mostly centered on telemedicine in the public sector, especially the eSanjeevani platform. Although this is consistent with the research question, it does not

consider the presence of private telemedicine services, which could be more relevant in urban or more affluent groups. This implies that the results can be less than the total telemedicine application and its effects on outpatient care in some areas.

Third, the OPD visit is categorized as non-urgent based on reported patterns of consultation as opposed to clinical triage information. There are also cases that are handled by telemedicine, which might also need a personal follow-up and this overlap cannot be exhaustively covered by secondary data sources.

The timeframe of the study also coincided with the COVID-19 pandemic and its consequences, whereby healthcare utilisation was affected by lockdowns, fear of infection, delayed care, and interruption of normal services. Even after the restrictions were lifted, telemedicine did not stop increasing, but these outside influencing factors constrain the possibilities of identifying change in OPD utilisation exclusively by expansion of telemedicine.

Lastly, differences in how data is reported by states and lack of specific pre-2020 telemedicine benchmarks limits comparisons over time. Future studies based on patient-level data, longer time frames, and mixed methods would be beneficial to understand the degree to which telemedicine replaces physical outpatient care.

Conclusion

This paper explored how the growth of telemedicine in India in the period 2020-2023 was linked with the variation in the number of non-urgent outpatient department (OPD) visits in government hospitals. The results indicate that with a significant rise in the usage of telemedicine- especially via the government-led eSanjeevani portal- physical visits to the OPD decreased significantly during the pandemic and then more gradually in the years that followed. The fact that the growth of teleconsultations is increasing and OPD growth is not, suggests that telemedicine has probably taken over some of the routine outpatient work, particularly in terms of follow-ups, chronic disease management, and minor health issues. Instead of just being a temporary measure amid the pandemic, telemedicine seems to have become an alternative to some outpatient services in the public health system in the long run.

Meanwhile, telemedicine has not eliminated the need to come to physical care, especially to complex, acute, or diagnostic cases. Its effect is to be, therefore, not total replacement but partial substitution. Continuous obstacles like digital literacy, disparities in infrastructure, and variations at state levels in adoption are still critical limitations.

In general, the results endorse the further implementation of telemedicine in the everyday provision of health services to the population in India. Telemedicine can alleviate OPD congestion, enhance the efficiency of the system, and make outpatient care more accessible to the populations that encounter barriers to frequent access to hospitals, provided it is implemented together with investments in digital access and primary care capacity.

Works Cited

Afrihyia, E., Forkuo, A. Y., Mustapha, A. Y., Chianumba, E. C., & Omotayo, O. (2025). Telemedicine and remote patient care: Expanding access and reducing inefficiencies in the healthcare system. World Scientific News, 205, 164–191.

Guleria, K. S., Patiyal, N., Negi, A. K., Kanwar, V., & Dinesh, K. (2020). Utilization of outpatient eSanjeevani National Teleconsultation Service during COVID-19 pandemic in a public healthcare institution in North India. Indian Journal of Pharmacy and Pharmacology, 7(4), 265–269.

Narwadiya, S. C., & Rao, D. R. (2025). Telemedicine in India: An impact analysis. Journal of Integrated Health Sciences, 13(1), 100004. https://doi.org/10.1016/j.inhs.2025.100004

Ramezani, R., Zhang, W., Cao, M., Bui, A., Petruse, A., Weldon, A., & Naeim, A. (2025). Remote patient monitoring for global emergencies: Case study in patients with COVID-19. JMIR Formative Research, 9, e66773.

Sood, S., Lal, K., Bhatia, M., Kapoor, G., Singh, S., Kaushish, R. K., … Verma, N. (2025). Adoption and utilization of India's eSanjeevani national telemedicine service. Oxford Open Digital Health, 3, oqaf025. https://doi.org/10.1093/oodh/oqaf025

Willis, V. C., Craig, K. J. T., Jabbarpour, Y., Scheufele, E. L., Arriaga, Y. E., Ajinkya, M., … Bazemore, A. (2022). Digital health interventions to enhance prevention in primary care: Scoping review. JMIR Medical Informatics, 10(1), e33518.

Das, P., Malhotra, S., Rahi, M., Chaturvedi, R., Chhibber-Goel, J., Anvikar, A., Shankar, H., Yadav, C. P., Meena, J., Tewari, S., Gopinath, S. V., Chhabra, R., & Sharma, A. (2021). Deployment of telemedicine as another mitigation tool during the COVID-19 pandemic in India. Public Health in Practice, 2, 100167.

Dastidar, B. G., Jani, A. R., & Suri, S. (2024). Reimagining India's national telemedicine service to improve access to care. Lancet Regional Health – Southeast Asia, 30, 100480.

Maroju, R. G., Choudhari, S. G., Shaikh, M. K., Borkar, S. K., & Mendhe, H. (2023). Role of telemedicine and digital technology in public health in India: A narrative review. Cureus, 15(3), e35986.

Naithani, C., Sood, S. P., & Agrahari, A. (2023). The Indian healthcare system turns to digital health: eSanjeevaniOPD as a national telemedicine service. Journal of Information Technology Teaching Cases, 13(1), 32–42.

Parthasarathi, A., George, T., Kalimuth, M. B., Jayasimha, S., Ullah, M. K., Patil, R., Nair, A., Pai, U., Inbarani, E., Jacob, A. G., Chandy, V. J., John, O., Sudarsanam, T. D., & Mahesh, P. A. (2024). Exploring the potential of telemedicine for improved primary healthcare in India: A comprehensive review. Lanc*et Regional Health – Southeast Asia*, *27*, 100431.

Rajkumar, E., Gopi, A., Joshi, A., Thomas, A. E., Arunima, N. M., Ramya, G. S., Kulkarni, P., Rahul, P., George, A. J., Romate, J., & Abraham, J. (2023). Applications, benefits and challenges of telehealth in India during COVID-19 pandemic and beyond: A systematic review. *BMC Health Services Research*, *23*(1), 7.

Shekhar, R., Sheikh, A. B., Upadhyay, S., & Upadhyay, S. (2021). Challenges in delivering primary care via telemedicine during COVID-19 pandemic in India: A review synthesis using systems approach. *Journal of Family Medicine and Primary Care*, *10*(9), 3208–3214.

From Separation to Synergy: Taking Advantage of Chemical and Biological Catalysts for Sustainable Synthesis By Rina Urabe

Abstract

Catalysts are important tools for scaling up many types of reactions to an industrial scale. Traditionally, catalysts have mainly been divided into chemical catalysts and biological catalysts, such as enzymes. This review analyses the transition from this divide towards a more modern, integrated approach by examining three key transformations: the reduction of unsaturated bonds, the reduction of nitrocompounds to amines, and terpene cyclisations. By analysing recent advancements in chemoenzymatic cascades and protein engineering, the review highlights how fusion systems outperform single-catalyst methods. Additionally, this review considers supramolecular catalysts, such as resorcinarene capsules, which mimic enzymatic microenvironments to stabilise reactive intermediates. These examples demonstrate how certain properties, such as the durability of traditional metal catalysts, can be combined with the selectivity of enzymes to create more sustainable methods. This analysis highlights that it is no longer efficient to focus solely on a single catalyst type and that the focus should shift towards synergising both biological and chemical systems to overcome the limitations of each. This review also explores the challenges of creating a "universal" catalyst, explaining why it is difficult to achieve a catalyst that is both hyperselective and widely applicable. Ultimately, this review demonstrates that combining the advantages of both catalyst types helps achieve better sustainability and efficiency.

Introduction

Catalysis is a fundamental aspect of modern chemistry across many fields, such as the fuel, pharmaceutical, and fragrance industries. By speeding up the reactions and producing stereoselective products without being consumed, it optimises the efficiency of almost all chemical reactions. The concept of catalysis first emerged in 1835, when Jöns Jakob Berzelius observed that certain species could increase the rate of a reaction whilst remaining unchanged. This discovery inspired other scientists to explore chemical reactivity, and over the next century, catalysis became essential to large-scale manufacturing and industrial chemistry.

For instance, the Haber-Bosch process developed in the early 1900s used an iron-based catalyst to produce ammonia from nitrogen and hydrogen. This method revolutionised agriculture by making nitrogen fertilisers more accessible to farmers. At the same time, advancements in cracking crude oil during petroleum refining emphasised the role of catalysis in modern industrial applications. Subsequently, the catalyst industry has grown rapidly, and catalysts are broadly classified into two categories: heterogeneous and homogeneous. Heterogeneous catalysts are catalysts that are in a different state from the reactants. They are mostly used in processes such as hydrogenation, oxidation, and cracking.

Homogeneous catalysts are catalysts that are in the same state as the reactants and form an intermediate species that is then regenerated in the final step. They are usually involved in transition-metal complexes or organic compounds.

Even though traditional catalytic systems have many advantages, they come with their own set of challenges. Many effective catalysts made from noble metals such as platinum, palladium, and rhodium are not only expensive and scarce but also costly to operate because they require high temperatures and pressures, thereby increasing energy consumption.

Consequently, chemical catalysts sometimes lack selectivity, leading to unwanted by-products that require additional purification. The importance of stereoselectivity was first brought up in the 1950's due to the thalidomide tragedy. Thalidomide was prescribed to pregnant women in the late 1950's as a treatment for morning sickness. However, the drug was sold as a racemic mixture of both the *R*-enantiomer, which acted as the sedative and the *S*-enantiomer, which was a potent teratogen which caused severe birth defects such as limb deformities. This public health crisis prompted the development of rigorous testing standards for pharmaceutical drugs, assessing both safety and efficacy in each enantiomeric form. These issues have prompted scientists to seek alternative catalytic systems that are more sustainable, selective, and efficient.

From the ancient Egyptians, biocatalysts such as the yeast used to ferment sugar in bread have been integrated into our everyday lives. In the late 19th century, Eduard Buchner demonstrated that cell-free extracts could convert sugars into ethanol. This was innovative at the time because it proved that biological activity was not limited to living cells but rather depended on specific protein molecules called enzymes. Since then, the biocatalytic industry has become a significant part of chemistry and biotechnology. Advancements in protein engineering have made it possible to design and modify enzymes to improve their stability and substrate range. Techniques for enzyme immobilisation have also been important, allowing enzymes to be attached to solid supports and reused multiple times, further improving their sustainability. Compared to traditional chemical catalysts, enzymes typically require less energy and produce fewer by-products, making them appealing for sustainable production. However, many biocatalysts have not yet been scaled up to industrial scale, and due to the selective nature of enzymes, they are not widely applicable to all reactions.

Recent developments have illustrated that chemical and biological catalysis should not be seen as opposing methods but rather as complementary approaches. Chemical catalysts have robust characteristics, function under extreme conditions and are used for a wide variety of reactions. In contrast, biocatalysts offer precision and require minimal energy use and waste. Combining both systems has led to hybrid or chemoenzymatic methods that take advantage of the strengths of each. Some examples of merged catalysts include transition-metal catalysis with biocatalysis, such as palladium with enzymes for reactions, or using light to drive reactions with a photocatalyst in combination with enzymes, leading to novel reactivities impossible with either catalyst alone. This trend is a part of a larger movement in modern chemistry towards atom economy, efficiency, and environmental responsibility.

The ultimate aim of research in catalysis is to create universal catalysts that can be used for a wide range of substrates with high selectivity, efficiency, and stability. Although scientists are far from this ideal, progress has been made in both the chemical and biocatalytic fields, which continue to bridge the gap to the universal catalysts. Additionally, breakthroughs in computational chemistry, machine learning, and structural biology, along with rapid developments in AI, have led to the integration of traditional empirical techniques with contemporary predictive models for improved catalyst design.

This literature review aims to demonstrate the advancements in both chemical and biocatalytic methods for three transformations, investigating their recent progress, benefits and limitations. The first transformation is the reduction of unsaturated bonds, an important step in producing flavour and fragrance compounds. The next transformation is the reduction of nitrocompounds to amines, an important process in the production of pharmaceuticals, agrochemicals, and fine chemicals. Lastly, terpene cyclisation will be explored by covering both the head-to-tail and tail-to-head mechanisms, including the conversion of nerol to limonene. Instead of determining which method is better, the focus will be on synergy, on how different types of catalysts can complement each other, and how this integration could potentially produce universal catalysts.

Main Text

Unsaturated bond reductions

The reduction of unsaturated bonds, such as the hydrogenation of alkenes to alkanes, is an important transformation in organic chemistry. Unsaturated bonds are highly reactive due to the presence of the pi electrons, so they are converted into saturated C-C bonds to form more stable molecules with modified physical and chemical properties. This process is used to control the molecular reactivity and also improve selectivity in multistep syntheses. The saturated products from these reductions are key intermediates for many industries, including fine chemicals in the fragrance industry and polymer production.

Olefine hydrogenation is an established method for reducing C=C bonds and has long been the traditional method for reducing alkenes. During this process, hydrogen gas reacts with the C=C bonds in the presence of a metal catalyst such as Nickel or Platinum, converting an alkene into an alkane. Heterogeneous catalysts have been the preferred catalysts in industry because they are reusable, easy to separate from products, and have high thermal stability. Therefore, this process was used for a variety of industries with applications including the synthesis of pharmaceuticals and the hardening of vegetable oils.

Although traditional olefin hydrogenation has been used in many industries, it has some drawbacks. A key challenge was the lack of chemoselectivity of the products formed in the process for molecules containing multiple reducible functional groups, resulting in over-reduction or unnecessary isomerisation of the double bonds and difficulty in reducing sterically hindered alkenes. Additionally, high-pressure operation poses safety risks and is

energy-intensive, and the use of noble-metal catalysts raises economic and environmental concerns.

Recently, nickel-based nanocatalysts have been used as a sustainable and efficient alternative to traditional catalytic methods. Studies have reported the development of supported nickel catalysts, including Ni/C materials derived from Ni-salen complexes and core-shell structures such as Ni@C-450. These systems demonstrate high selectivity, enabling the reduction of a C=C bond in the presence of other reducible functional groups such as aldehydes, ketones, esters, nitriles, and halides. Furthermore, these improved methods require only mild conditions, and the catalyst can be reused and recovered over and over again. Another method involves generating Nickel nanoparticles *in situ*, which allows for efficient hydrogenation of sterically hindered molecules whilst maintaining broad functional-group tolerance. Therefore, this environmentally friendly and cost-effective method is a modern alternative to traditional noble-metal catalysts.

Biocatalysis offers a novel synthetic pathway for high-value molecules such as raspberry ketones (Fig. 1), which are in high demand in the fragrance and food industries. A major issue in the past has been that people often look for raspberry ketone in its natural form rather than the synthetic version. Therefore, many scientists have developed biocatalytic methods that allow raspberry ketones to remain in their natural form, driving up prices. Early studies demonstrated that yeast extracts could catalyse the reduction of C=C bonds using the biological cofactor NADPH as the hydrogen source. Additional screening and research on microorganisms found that several other fungi and bacteria were also able to produce raspberry ketones, including *Mucor Subtilissimus,* which showed particularly promising efficiency, produced few by-products, and reacted rapidly. These biocatalytic systems offer high selectivity, mild conditions and an environmentally friendly approach to reducing unsaturated bonds. However, with high selectivity comes the issue that scaling up biocatalysts for industry is difficult, as the catalysts can be expensive to produce and require a specific temperature to avoid denaturation. Additionally, biocatalysts tend to hyperspecialise, so they cannot be used across a wide range of compounds, limiting their flexibility and efficiency in industry.

Figure 1. Reduction of (*E*)-4-(4- hydroxyphenyl)but-3-en-2-one to raspberry ketone.

To bridge the gap between chemical durability and biological precision, recent research focuses on chemoenzymatic cascades. For instance, a recent study demonstrated the use of a one-pot chemoenzymatic cascade for the asymmetric synthesis of a key chiral intermediate in the production of an antiepileptic drug, brivaracetam. This method combined biocatalysis with

ene-reductases and a protein-engineered alcohol dehydrogenase (ADH) to shape the molecule, and chemocatalysis was represented by the chemical cyclisation step, forming a ring structure. This method achieves over 98% enantiomeric excess but still lacks industrial scalability, as producing enzymes that use NADPH cofactors to supply hydrogen is expensive and NADPH is not regenerated.

Another example of chemoenzymatic synergy combines palladium-catalysed Suzuki-Miyaura cross-coupling with whole-cell biocatalysis to achieve the asymmetric reduction of carbonyl groups, producing chiral alcohols (Fig. 2). This method uses a metal-catalysed step to produce the molecular framework, and whole-cell enzymes then catalyse the enantioselective reduction in the cell. Additionally, the cell membrane serves as an additional barrier that prevents harmful organic solvents from entering the cell. This method of integrating transition-metal catalysis with enzymes improves enantioselectivity and product yield, demonstrating the potential of chemoenzymatic systems for efficient and sustainable synthesis of chiral compounds.

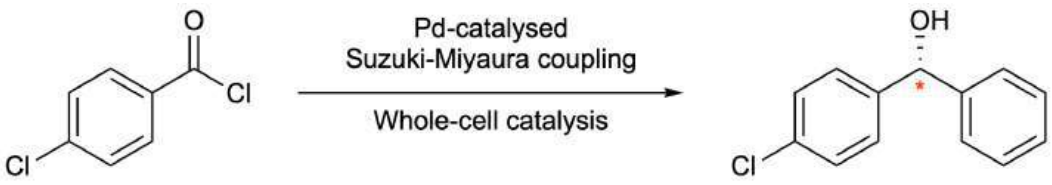


Figure 2. Chemoenzymatic cascade combining Suzuki-Miyaura cross-coupling with whole-cell biocatalysis to achieve the asymmetric reduction of carbonyl groups, producing chiral alcohols.

Overall, the development of these methods balances cost, efficiency, sustainability and selectivity. Traditional heterogeneous hydrogenation remains the standard method for reducing unsaturated bonds due to its scalability and reliability, although new alternatives are still being investigated to improve selectivity and reduce environmental impact. Modern nickel-based nanocatalysts improve the chemoselectivity and use more cost-effective metals, whilst advances in biocatalysts have led to enhanced selectivity and the production of high-value compounds under environmentally sustainable temperatures. However, these biocatalysts have limited industrial applications due to their limited scope, reliability, and cost. The most promising path forward lies in integrating transition-metal catalysis with enzymes to create efficient, sustainable, and highly selective "hybrid" systems for the future.

Reductions of nitrocompounds to amines

Producing amines from nitrocompounds such as nitroalkanes and nitroarenes is a core transformation. Nitro compounds usually contain nitrogen in a high oxidation state, such as +3 or +5; therefore, reducing them to amines yields a more reactive, versatile amine. Amines are used as building blocks in a variety of industries, including as active pharmaceutical ingredients (APIs) in pharmaceuticals, key intermediates for agrochemicals and polyurethane monomers, and base reactants for forming dyes and pigments.

The Bechamp Reduction in the mid-19th century was one of the first catalytic methods developed for reducing nitrocompounds to anilines using Fe/HCl. This technique was important in the pigment industry as the aniline form was later used to create synthetic dyes such as indigo, aniline black, and methylene blue. Additionally, this method was also used to undergo the large-scale synthesis of 4,4'- methylenedianiline (MDA), which was used for polyurethane production. This process was a major breakthrough, as it replaced the previous Sn/HCl reactions with Fe, which was widely available, enabling the reaction to be scaled up for industrial applications.

However, the reaction produced many by-products that were harmful to the environment. For instance, more iron oxide formed than aniline, and it was difficult to dispose of, leading to significant waste management issues. The lack of selectivity in this reaction produced azobenzene as a by-product, which meant this process would require additional time-intensive and expensive steps to separate the anilines. Not only that, but the reaction was carried out under acidic conditions, which corroded equipment and made wastewater treatment difficult.

Despite these setbacks, it was crucial in encouraging industrial catalysis research as well as synthetic dye development. This led to early metal hydrogenation in the 20th century, with the introduction of Pd, Ni, Ru, and Pt catalysts. For instance, Raney Nickel, a Ni/Al alloy, became a widely applicable heterogeneous catalyst developed by Murray Raney. The Ni/Al alloy was an effective catalyst due to its large surface area from the finely divided Ni. The Pd/C also became a versatile catalyst, as it was compatible with a wide variety of functional groups and could operate under mild conditions. Industrial hydrogenations using Pd and Ni were used to produce aniline derivatives and intermediates for nylon synthesis in the textile industry. However, this process was still energy intensive as it required high pressures (10-100 bar) and temperatures (100-200C). Additionally, although selectivity improved, these catalysts could still reduce not only the nitro group but also C=O and C=C bonds, as well as dehalogenate aromatic substrates.

This laid the foundations for modern heterogeneous catalysts for reducing nitrocompounds to amines by spurring research into metal-support interactions, bimetallic systems, and selective hydrogenation techniques. The production of catalysts, such as bimetallic catalysts such as Pt-Fe/TiO_2 and Rh-Fe/C, and modified catalysts such as Pt or Pd with Zn, Pb, Bi, Sn, which control selectivity and avoid unwanted chemical reactions by regulating metal-support interactions and controlling electron density for better chemoselectivity. The metal supports, such as Pt-Fe/TiO_2, increase the surface area, improve metal dispersion, and facilitate mass and energy transfer to enable a uniform reaction. This is a big change compared to standard Pt, Pd, or Ni catalysts, which cannot reduce complex nitro compounds bearing double bonds or halogens without producing harmful by-products.

Another advancement was the shift towards noble-metal catalysts, such as Au and Ag nanoparticles. For instance, Corma's group found that Au nanoparticles on TiO_2 exhibited excellent selectivity for the reduction of nitroarenes, a selectivity rarely matched by other noble metal catalysts due to the preferential adsorption offered by the Au nanoparticles on the TiO_2 surface. Thus, allowing for the reduction of more complex nitroarenes that have multiple

functional groups. This discovery led to further research on Au and Ag catalysts and demonstrated that selectivity can arise from varying the support-metal interaction, in addition to changing the metals. This catalyst seems promising for selective hydrogenation on an industrial scale; however, this catalyst has not been widely applied in large-scale industrial processes yet.

Although these processes offer speed and efficiency, they come with some setbacks. For instance, the scarcity of noble metals such as Pt increases costs due to the challenging, energy-intensive processes of recycling and recovering these metals. Additionally, reliance on rare metals limits sustainability, as they are nonrenewable resources. Also, the toxic solvents (e.g., toluene) used in these processes pose health hazards to workers, and wastewater often contains trace metals that pollute the surrounding environment.

Furthermore, trace impurities such as nitrogen and sulfur can adsorb onto the metal surface, blocking active sites, thereby reducing the reaction rate, lowering selectivity, and, in severe cases, permanently deactivating the catalyst, requiring more catalyst replacements and thus increasing the cost and waste produced.

Biocatalysts such as nitroreductases and flavoproteins eliminate the need for noble metals in the reduction of nitrocompounds to amines. Enzyme active sites are especially regio- and chemoselective, targeting the nitro group, thereby reducing unwanted by-products in the reduction of nitrocompounds. Protein engineering has been used to produce nitroreductases with a higher substrate allowance, even for heteroaromatic compounds. Another advancement is reduced catalyst poisoning, as biocatalysts are naturally resistant to sulfur poisoning because they require mild aqueous conditions, and encapsulating sensitive catalysts within polymers can prevent contamination of the nitroreductases.

Recent studies suggest that hydrogenation of nitro groups on palladium on carbon (Pd/C) may proceed by first dissociating hydrogen on palladium sites (Fig. 3). This produces electrons that then travel to the carbon support, where they reduce the nitro compound. However, when Pd nanoparticles are used with hydrogen, unwanted side reactions can happen. These include hydrogenation of other double bonds, dehalogenation, and poisoning of the catalyst if sulfur is present in the molecule.

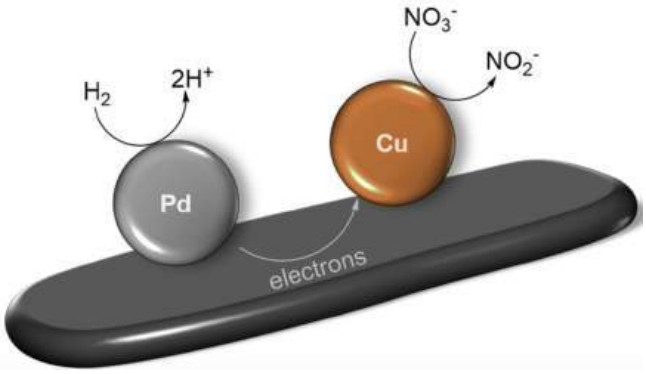


Figure 3. Schematic representation of a heterogeneous supported metal alloy in which H_2 oxidation occurs on Pd sites and nitrate reduction on Cu sites. Reproduced with permission from *Chem. Commun.* (2024) 60 (93): 13667–13677.

Inspired by this recent study of the mechanism of nitro group reduction of Pd/C catalyst, researchers suggested that hydrogenase on carbon can be used as a selective catalyst to activate hydrogen and carry out palladium-free electrochemical hydrogenation of nitro compounds (Fig. 4). Hydrogenase enzymes avoid the side reactions associated with electrochemical hydrogenation by splitting hydrogen at a hidden active site that large organic molecules cannot reach. The electrons produced are moved through a chain of iron-sulfur clusters to the protein surface.

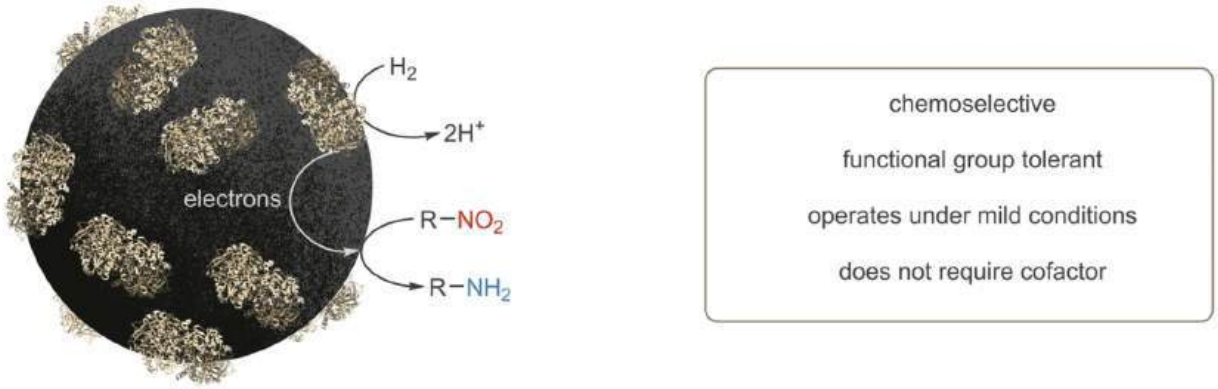


Figure 4. Hydrogenase enzyme (gold) immobilised on the surface of carbon black particles as a catalyst for efficient chemoselective synthesis of amines. Reproduced with permission from *Nature Communications*, vol. 15, no. 1, 24 Aug. 2024.

However, there are still setbacks in biocatalysis. For instance, biocatalysts are so selective that each nitroreductase works only with specific substrates; therefore, biocatalysis is not universal for all reductions of nitrocompounds. Additionally, these processes have not been scaled up to industrial scale yet, so their impact has not yet been observed.

The reduction of 2-methyl-5-nitro-pyridine to 2-amino-5-methyl-pyridine is an important transformation and is an essential building block molecule in the pharmaceutical industry. Traditional metal catalysis with these molecules is difficult because catalyst poisoning by nitrogen-containing heterocycles reduces efficiency and limits applicability. Additionally, biocatalysis is difficult because natural enzymes are slow and denature under substrate inhibition. Therefore, researchers combined both methods to produce an optimal catalyst.

The researchers used rational engineering to create variants of NR5 from the NR4 framework, which was already able to catalyse the reduction of nitrocompounds. They used a structural homology model (a 3D computer map of the protein) to identify five specific amino acids in the active site that were most likely causing the substrate inhibition. The NR5 enzyme is key because it proved that enzymes could be used on a larger scale (up to 50 g) to catalyse nitroreductions. V_2O_5 is used together with the enzyme as a disproportionation agent to drive the reaction to achieve 95% yield of the final product and complete the reduction of the hydroxylamine to a primary amine (Fig. 5).

This method has been shown to be effective because the enzyme exhibits high chemoselectivity; unlike traditional metal catalysts, it reduces only the nitro group, leaving other functional groups, such as double bonds or halogens, unaffected. Additionally, the NR5 enzyme has a high concentration tolerance but only requires mild reaction conditions, ~30 °C. However, although this is significant progress, several limitations remain. For instance, V_2O_5 is unsustainable and toxic to the environment. Furthermore, the enzymes require sugar to keep producing life, producing a lot of biowaste; for every gram of amine produced, a gram of gluconic acid byproduct is made.

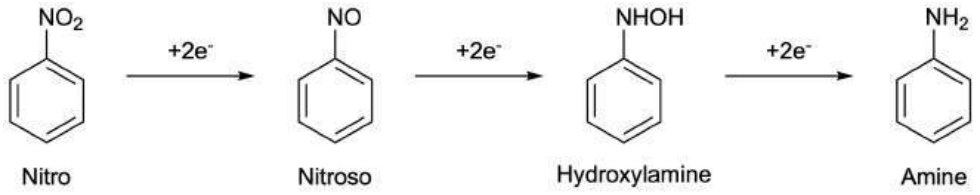


Figure 5. Step-by-step 6-e^- reduction of nitrobenzene to aniline via nitroso and hydroxylamine intermediates.

In addition to producing more by-products, it requires more materials to produce the enzyme itself, which are expensive because they are cofactor-dependent and require NADPH, which is more costly (thousands of USD per kilo) than the cost to extract metals like nickel. This is because producing the NADPH requires two enzymes, Glucose Dehydrogenase (GDH) [the power generator] and Nitroreductase (NTR) [the main enzyme]. Therefore, even with regeneration, this method is not efficient because both the initial cofactor and the secondary enzyme (GDH) are expensive.

However, its greatest limitation is that the catalytic process is too specialised and not as universal as palladium's. Although there are setbacks to this method, it represents an overall step in the right direction and demonstrates the concept of synergy in combining chemical and biocatalytic approaches. Therefore, in the future, more engineered hybrid chemoenzymatic systems can be utilised to link robust metal catalysts with enzymes, using biological conditions designed so that each catalyst performs its individual role. For instance, metals that handle redox and nitroreductases focus on filtering molecules to avoid toxic by-products. This flexibility would allow for customised catalytic systems that can be adapted to a wide range of applications.

Terpene cyclisations

Terpene cyclisation is an important transformation in both organic chemistry and natural biosynthesis. In these transformations, linear isoprenoid precursors, such as geranyl derivatives, are converted into complex cyclic structures via carbocationic cascade reactions. These cascades generate many C-C bonds and stereocenters simultaneously. This transformation is commonly used to construct the core frameworks of mono-, sesqui-, and diterpenes, significantly reducing the number of steps in the transformation for complex molecules whilst increasing the overall yield of the reaction.

Cyclic terpenes are used in a wide variety of industries, but are especially useful in the fragrance, fine chemical, and pharmaceutical industries. Their structural range is directly linked to the versatility of the cyclisation pathways. In nature, terpene cyclases convert simple acyclic diphosphates into complex molecules with high selectivity under mild conditions. The aim is to replicate this selectivity in the lab to slow the depletion of natural resources.

Typically, terpene cyclisation is carried out using strong Brønsted-Lowry acids to generate a carbocation. The protonation generates an allylic carbocation, which then undergoes an intramolecular attack by the neighbouring double bonds to form cyclic products. This approach uses biomimicry to mirror the tail-to-head cyclisation observed in nature.

However, in solution, it is difficult to control these reactions because the highly reactive carbocation intermediates can undergo alternative mechanisms, such as elimination or polymerisation. The transformation of simple monoterpene alcohols into complex terpene products often leads to a mixture of unwanted by-products. Therefore, this traditional method incurs additional costs because separating the useful products from the by-products requires more energy and is less environmentally friendly. Additionally, functional-group tolerance is limited in this process, so there are constraints on which molecules can be used in the traditional method. Open solution systems lack the precise substrate preorganisation and intermediate stabilisation that enzyme active sites have.

Another biocatalytic approach to terpene cyclisation is to protein engineer terpene enzymes to produce high-value terpenes. Protein engineering enables directed evolution, allowing enzymes to have a wider substrate scope and improved catalytic efficiency whilst maintaining stereoselectivity and near-neutral reaction conditions.

In contrast, enzymatic systems carry out these transformations in confined hydrophobic spaces. In one study, it was demonstrated that an enzyme-like selectivity can be mimicked using a self-assembled resorcinarene capsule. This capsule provides an aromatic cavity that stabilises carbocation intermediates through cation-π interactions and controls access of external nucleophiles. Within these spaces, the study showed that monoterpene alcohols such as nerol and geraniol can undergo stereoselective cyclisation to products including α-terpinene and eucalyptol. This work shows that selectivity can be strongly influenced by spatial confinement and by weak Brønsted acidity relative to bulk-solution reactions.

Whilst traditional chemical methods use harsh Brønsted acids that lack chemo-selectivity, recent advancements have focused on mimicking the microenvironment of natural cyclases. In nature, enzymes use complex hydrophobic pockets to isolate reactive intermediates. Zhang and Tiefenbacher (2015) have demonstrated that this chemoselectivity can be recreated using a hexameric resorcinarene capsule. This supramolecular structure self-assembles in apolar solvents, forming an eight-water-molecule and six-resorcinarene unit complex that creates a hydrophobic cavity that prevents premature reactions with external nucleophiles.

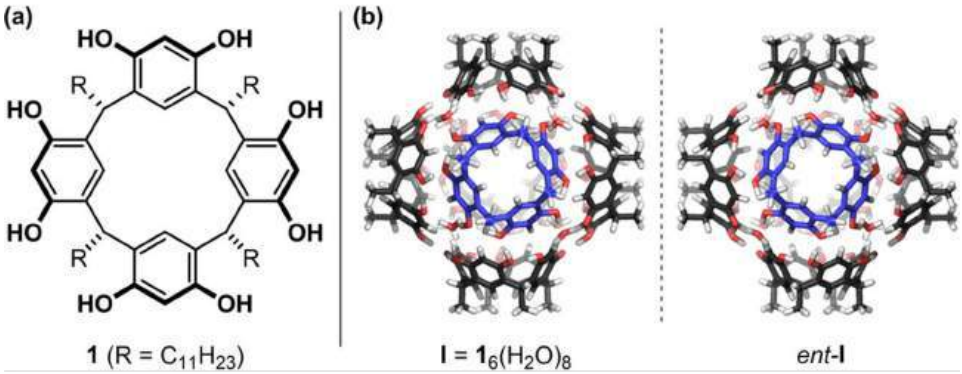


Figure 6. Self‑assembly of resorcin[4]arene 1 (a) into the chiral but racemic hexameric capsule I (b). Reproduced with permission from *Angew. Chem. Int. Ed.* **2022**, *61*, e202203384.

The resorcinarene capsule also mimics the amino acid residues in natural enzymes, with electron-rich aromatic walls that stabilise these cations through cation-pi interactions. Additionally, this method has a high atom economy because more cations yield the desired product and less by-product is formed. It was also shown that nerol cyclisation inside a derivatised resorcinarene capsule can happen in an enantioselective manner, leading to the formation of *a*-tepreniol (60% *ee*) and limonene (70% *ee*) (Fig. 7).

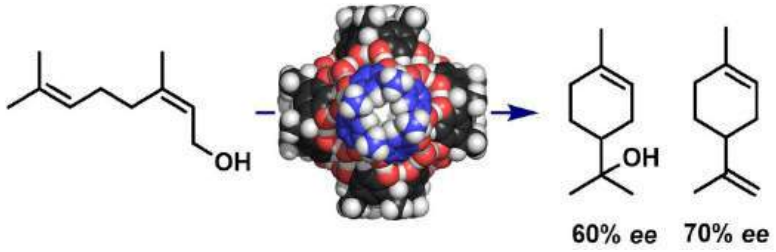


Figure 7. Enantioselective cyclisation of nerol inside the resorcinarene capsule leads to formation of *a*-tepreniol (60% *ee*) and limonene (70% *ee*). Reproduced with permission from *Angew. Chem. Int. Ed.* **2022**, *61*, e202203384.

However, there are still some limitations to this method. Although the capsule prevents some reactivity, competing side reactions still occur, such as dimerisation and alkylation of the capsule's own phenol groups.

Terpene synthases catalyse the conversion of linear prenyl pyrophosphate substrates into cyclic terpene products through reactive carbocation intermediates. The active site geometry influences the reaction to form specific products through the steric arrangement of the substrate and the stabilisation of intermediates. The Johnson and Allemann paper demonstrates recent advancements in the engineering of these enzymes through subtle point mutations within the active site that could alter product distribution.

For example, the sesquiterpene synthase delta-cadinene synthase was engineered to produce up to 90% of the hydroxylated alcohol, germacradien-4-ol, instead by increasing the volume of the active-site cavity to allow nucleophilic water access to the final carbocation intermediate. This is important because it was unachievable with either Brønsted acid catalysis or

supramolecular capsule systems. The research suggested that terpene synthases can be rationally engineered using a combination of X-ray crystallography, molecular dynamics simulations, and targeted mutagenesis. This was achieved by controlling water at the active site, using unnatural substrate analogues, and improving terpene production via biphasic flow systems. Non-natural substrate analogues were used to expand the chemical space accessible to terpene synthases.

However, the catalytic efficiency of terpene synthases remains a significant challenge, thereby limiting their industrial viability. The typical catalytic turnover numbers for terpene synthases are much lower than for other enzymes due to the low solubility of terpene products, which slows their release from the enzyme. To fix this, a biphasic continuous-flow system was developed using pentane, which doubled catalytic activity and yielded greater than 50% for several sesquiterpene synthases. Although this was successful on a lab scale, it remains unclear whether it can compete with chemical synthesis on a tonne scale.

Another important development in terpene cyclisation is the use of chemoenzymatic systems, which combine chemical and enzymatic catalysis to overcome the limitations of each approach. These hybrid strategies combine the strengths of both methods, such as the scalability and flexibility of chemical synthesis alongside the selectivity of these enzymes.

One example of this synergy is shown through the one-pot enzymatic synthesis of merochlorin A and B. Both intermediates are chemically synthesised, and then the enzymatic steps are combined to reconstruct a full biosynthetic pathway. In this study, a polyketide-terpene hybrid intermediate was first formed via biocatalysis with enzymes, followed by further chlorination and terpenecyclisation catalysed by a single vanadium-dependent haloperoxidase. This transformation is important because it shows that complex terpene cyclisation reactions, which usually require multiple steps in synthetic chemistry, can be achieved in a highly selective manner using minimal enzymes.

Additionally, the study demonstrated how natural systems achieve high structural diversity with limited catalytic components by using only four dedicated enzymes to generate structurally complex meroterpenoids. This differs from traditional chemical synthesis, in which multiple protecting-group strategies and step-by-step transformations are required. However, the process still relied on chemically synthesised intermediates such as pre-merochlorin, suggesting that chemical synthesis remains important for accessing certain substrates that are not easily produced via biocatalysis. The advantage to this method is that chemical synthesis can be used to construct complex or non-natural intermediates, which can then be selectively transformed by enzymes. Although this approach is still limited by enzyme availability, substrate compatibility, and relatively low catalytic turnover compared to industrial chemical catalysts.

Another example of synergy is the modular chemoenzymatic synthesis of terpene analogues, in which simple alcohol precursors such as prenol and isoprenol are first made chemically accessible, then enzymatically converted into diphosphate intermediates, and finally cyclised into terpene products. This system uses a variety of enzymes, including kinases, prenyl transferases, and terpene synthases, to generate both natural and non-natural terpenes.

The advantages of this method are that it avoids traditional chemical methods for diphosphate synthesis, which are usually low-yielding and require extra purification steps, making them industrially infeasible. Instead, the chemoenzymatic pathway allows for quantitative yields in the phosphorylation steps and yields greater than 80% for important intermediates such as farnesyl. This shows how enzymatic steps can improve efficiency and reduce byproducts compared with purely chemical routes.

Furthermore, this system enables access to terpene structures from non-natural substrate analogues that are not found in nature. For instance, modified diphosphates were cyclised into unnatural sesquiterpenes with altered properties, such as switching from insect repellents to attractants. This allows enzymes to have a more diverse range of substrates, thereby enabling them to catalyse a wider variety of terpenes. However, this system requires careful control of enzyme combinations and reaction conditions, such as magnesium ion concentrations and ATP recycling systems, making it difficult to scale up from the lab to the industrial scale.

Comparing these two systems, both display the advantages of combining chemocatalysis with enzymes, but have different applications. The merochlorin study focuses on reconstructing complex natural product biosynthesis with as few enzymes as possible, prioritising efficiency and biomimicry. The modular chemoenzymatic system focuses on expanding chemical diversity and producing the synthesis of non-natural terpenes for versatility.

Conclusion and Outlook

In conclusion, the concept of catalysis and its uses have evolved from Berzelius' 1835 observations to the chemoenzymatic cascades used today. By looking at three different transformations, the unsaturated bond reductions, reduction of nitro compounds to amines, and terpene cyclisation, it demonstrates that the split between chemical and biological methods has faded and instead has been joined together to combine the advantages of each side. Whilst traditional chemical methods such as metal catalysts can be scaled to industry and are reliable, they are often harmful to the environment and have poor selectivity. In contrast, biological methods such as enzyme-based approaches are sustainable and offer excellent selectivity but are costly and hypersensitive. This demonstrates that the "synergy" of these two fields offers the best solution in the future. By combining the durability of transition-metal catalysts with the stereoselectivity of biocatalysts, researchers have created hybrid systems that achieve higher yields and efficiencies than either method alone. However, this challenges the idea of a "universal catalyst" as it could be paradoxical. This is because it is difficult to balance a catalyst's promiscuity and selectivity. A catalyst that works for many different types of substrates usually lacks the selectivity required to produce a singular product. Therefore, the progress of future catalytic methods should not focus on finding a single universal catalyst, but on developing efficient methods to identify individual catalytic cascades suited to different functional groups. Nowadays, with the rise of computational modelling and AI, it could aid the development of more optimal catalysts with better atom economy, efficiency and sustainability.

Works Cited

Fan, Xin-Yue, et al. “Applications of Ene-Reductases in the Synthesis of Flavors and Fragrances.” Journal of Agricultural and Food Chemistry, vol. 72, no. 33, 5 July 2024, pp. 18305–18320, https://doi.org/10.1021/acs.jafc.4c02897.

Feng, Jiacheng, et al. “One-Pot Chemo-Enzymatic Cascades Using Ene-Reductase and Alcohol Dehydrogenase to Produce a Brivaracetam Chiral Precursor.” Process Biochemistry, vol. 153, 2 Apr. 2025, pp. 267–275, www.sciencedirect.com/science/article/abs/pii/S1359511325001060, https://doi.org/10.1016/j.procbio.2025.04.001.

Papadopoulou, Athena, et al. “Development of an Ene Reductase-Based Biocatalytic Process for the Production of Flavor Compounds.” Organic Process Research & Development, vol. 26, no. 7, 14 June 2022, pp. 2102–2110, https://doi.org/10.1021/acs.oprd.2c00096. Accessed 13 Mar. 2024.

Wang, Kewei, et al. “Acid and Base Coexisted Heterogeneous Catalysts Supported on Hypercrosslinked Polymers for One-Pot Cascade Reactions.” Journal of Catalysis, vol. 348, 2 Apr. 2017, pp. 168–176, www.sciencedirect.com/science/article/abs/pii/S0021951717300581, https://doi.org/10.1016/j.jcat.2017.02.024.

Yin, Hang, et al. “Coupling Metal and Whole-Cell Catalysis to Synthesize Chiral Alcohols.” Bioresources and Bioprocessing, vol. 9, no. 1, 8 July 2022, https://doi.org/10.1186/s40643-022-00560-0. Accessed 15 Apr. 2024.

Bornadel, Amin, et al. “Process Development and Protein Engineering Enhanced Nitroreductase-Catalyzed Reduction of 2-Methyl-5-Nitro-Pyridine.” Organic Process Research & Development, vol. 25, no. 3, 10 Dec. 2020, pp. 648–653, https://doi.org/10.1021/acs.oprd.0c00464. Accessed 23 Mar. 2026.

Formenti, Dario, et al. “Reduction of Nitro Compounds Using 3d-Non-Noble Metal Catalysts.” Chemical Reviews, vol. 119, no. 4, 5 Dec. 2018, pp. 2611–2680, https://doi.org/10.1021/acs.chemrev.8b00547.

Sokolova, Daria, and Kylie A. Vincent. “Exploiting Hydrogenases for Biocatalytic Hydrogenations.” Chemical Communications, vol. 60, no. 93, Royal Society of Chemistry (RSC), 2024, pp. 13667–77, https://doi.org/10.1039/d4cc04525d.

Sokolova, Daria, et al. “Selective Hydrogenation of Nitro Compounds to Amines by Coupled Redox Reactions over a Heterogeneous Biocatalyst.” Nature Communications, vol. 15, no. 1, 24 Aug. 2024, https://doi.org/10.1038/s41467-024-51531-2.

Sokolova, Daria, et al. “Enantioselective Tail‐To‐Head Terpene Cyclizations by Optically Active Hexameric Resorcin[4]Arene Capsule Derivatives.” Angewandte Chemie International Edition, vol. 61, no. 25, 22 Apr. 2022, https://doi.org/10.1002/anie.202203384.

Trost, Barry M., and Chang Min. “Total Synthesis of Terpenes via Palladium-Catalysed Cyclization Strategy.” Nature Chemistry, vol. 12, no. 6, 1 June 2020, pp. 568–573, pubmed.ncbi.nlm.nih.gov/32231261/, https://doi.org/10.1038/s41557-020-0439-y.

Zhang, Q., and K. Tiefenbacher. “Terpene Cyclization Catalysed inside a Self-Assembled Cavity.” Nature Chemistry, vol. 7, no. 3, 16 Feb. 2015, pp. 197–202, https://doi.org/10.1038/nchem.2181.

Gambacorta, Guido, et al. “A Comprehensive Review of Flow Chemistry Techniques Tailored to the Flavours and Fragrances Industries.” Beilstein Journal of Organic Chemistry, vol. 17, 18 May 2021, pp. 1181–1312, www.ncbi.nlm.nih.gov/pmc/articles/PMC8182698/,

Jiang, Yanlong, and Hans Renata. “Modular Chemoenzymatic Synthesis of Ten Fusicoccane Diterpenoids.” Nature Chemistry, vol. 16, no. 9, 6 May 2024, pp. 1531–1538, www.nature.com/articles/s41557-024-01533-w, https://doi.org/10.1038/s41557-024-01533-w.

Johnson, Luke A, et al. “Modular Chemoenzymatic Synthesis of Terpenes and Their Analogues.” Angewandte Chemie International Edition, vol. 59, no. 22, 27 Feb. 2020, pp. 8486–8490, https://doi.org/10.1002/anie.202001744. Accessed 5 Feb. 2025.

Teufel, Robin, et al. “One-Pot Enzymatic Synthesis of Merochlorin A and B.” Angewandte Chemie (International Ed. In English), vol. 53, no. 41, 6 Oct. 2014, pp. 11019–11022, pubmed.ncbi.nlm.nih.gov/25115835/#:~:text=One-pot%20enzymatic%20synthesis%20of%20merochlorin%20A%20and%20B, https://doi.org/10.1002/anie.201405694. Accessed 27 Feb. 2022.

An Anatomy of Allegiance: The Dual Nature of Fidelity in the East By Hrishik Malhotra

Abstract

In the sixteenth and seventeenth centuries, Tokugawa Japan and the Mughal Empire faced a similar political problem: how to channel the strength of armed elites into state service while preventing them from destabilizing the regime. Historians have long studied samurai ethics and Mughal military institutions separately - the former through Neo-Confucian *Bushidō* narratives and the latter through the *mansabdari* system - but there has been little direct comparison. This essay addresses that gap by analyzing how each polity defined loyalty and duty in practical terms. Reading early modern sources side by side,[1] It argues that Tokugawa rulers reframed martial language into household laws and moral training, whereas the Mughal state embedded military obligation in a formal system of rank, inspection, and imperial favor. In both cases, the objective was to make warrior power predictable and accountable through internalized discipline in Tokugawa Japan and external administration in the Mughal Empire.

This argument is developed in four parts. First, it defines what is meant by a "warrior ethic" in this context: the set of norms and regulations that link martial virtue to legitimate rule. Second, it sketches the origins of samurai and *sipahi* identities before they were codified. Third, it examines how each regime legally institutionalized those norms (for example, Tokugawa edicts versus Mughal *mansabdari* regulations). Fourth, it considers how loyalty and discipline actually worked in practice and under stress. Throughout, I draw on primary sources - the *Buke Shohatto* (1615 Tokugawa code), *Hagakure* (samurai manual), the *Ain-i-Akbari* (Abu'l-Fazl's administrative compendium), and imperial *farmans* - as well as secondary scholarship. By reading these materials together, the essay shows that "*Bushidō*" and the Mughal military code were not fixed cultural inheritances but evolving tools of governance.

Keywords Military History, Samurai, Sipahi, Tokugawa Japan, Mughal Empire, Military Code

Warrior Ethics

In this study a "warrior ethic" is not a timeless set of virtues but the concrete expectations imposed on warriors by their state. In both polities, martial values appear primarily in prescriptive texts, legal codes, household regulations, and didactic writings, rather than in abstract philosophical treatises. For example, the Tokugawa *Buke Shohatto* (Laws for the Military Houses, 1615) addresses samurai behavior directly[2]. Article 1 counsels that "war is a

[1] Vaporis, Constantine Nomikos. *Breaking Barriers: Travel and the State in Early Modern Japan*. Harvard University Asia Center, 1994 pp. 27–34.

[2] "Buke Shohatto" ["Laws for the Military Houses"]. *Sources of Japanese Tradition*, edited by Wm. Theodore de Bary et al., vol. 1, Columbia UP, 2001, pp. 393–95.

curse" and even in peacetime one must "train yourself and be prepared[3]," while Article 2 forbids "group drinking and wild parties", equating martial virtue with temperance. These rules show that Tokugawa discipline was defined as regulated comportment - study, sobriety, and hierarchy - rather than battlefield glory.

Likewise, in Mughal India Abu'l-Fazl's *Ain-i-Akbari* translates military service into bureaucratic terms. Each noble's *mansab* (rank) prescribed exactly how many cavalry he must maintain and what quality of horses to keep. To enforce this, the state instituted detailed inspection rules. As one account notes, officers "were subject to detailed regulations regarding inspection of troops, branding and registration of horses and troopers, all designed to link numerical rank to real military strength and to prevent fraudulent musters[4]." In other words, a *sipahi's* (a Mughal foot soldier) loyalty was expressed through paperwork and muster rolls as much as through martial valor. In both Japan and India, then, the warrior ethic was an ensemble of rules and rituals that made loyalty and discipline legible to the government, rather than an individual moral creed written in scrolls of code.

Origins and Genealogies

The pre-modern origins of samurai and *sipahi* differed sharply. Samurai identity was forged in the incessant warfare of the Sengoku era (c.1467 - 1600). Provincial lords built armies of hereditary retainers, and battles like the five engagements at Kawanakajima (1553 - 1564) were contests between clan blocs[5]. In this context loyalty meant service to one's immediate lord, which was a relationship often renegotiated whenever a lord could no longer protect his vassals. There was little unified ideology beyond clan honor. After Tokugawa Ieyasu's victories at Sekigahara (1600) and Osaka (1615), long peace set in. Tokugawa intellectuals then reinterpreted the samurai past. Neo-Confucian thinkers such as Yamaga Sōkō and Daidōji Yūzan drew on concepts of propriety and virtue to give the warrior class a new justification. For example, in *The Way of the Samurai (Shidō)* Yamaga Sōkō insisted that in an era of peace the samurai "devote himself to duty above all" by cultivating virtue and serving his lord faithfully[6]. He argued that a warrior born into an aristocratic family must examine his calling and reinforce loyalty and fidelity in all relationships (master-retainer, family, friends). In effect, Tokugawa ideology retrospective historicized samurai as moral exemplars rather than active warriors, stabilizing their status while divesting them of independent military power.

By contrast, the Mughal *sipahi* (cavalry officer, usually a *mansabdar*) traced its identity to Timurid and Central Asian traditions of mobile cavalry warfare. Babur's memoirs recount hit-and-run tactics[7] and personal allegiance to the emperor, where speed and horsemanship were prized above all. Under Akbar, these martial norms were subsumed into an integrated imperial

---

3 "Buke Shohatto," p. 394.

4 Abu'l Fazl ibn Mubarak. *The Ain-i-Akbari*. Translated by H. Blochmann and H. S. Jarrett, Asiatic Society of Bengal, 1873 - 1907. 3 vols, pp. 1:252–255.

5 Turnbull, Stephen. *Samurai Warfare*. Cassell, 1996, pp. 45–52.

6 Yamaga, Sokō. *The Way of the Samurai*. Translated by Thomas Cleary, Shambhala, 2007, pp. 12–18.

7 Babur. *Baburnama*. Translated by Wheeler M. Thackston, Modern Library, 2002. pp 178–183.

theory of kingship. Abu'l-Fazl portrays Akbar as the "fountain of honor[8]" whose *amanat* (divine trust and authority) bestows rank and requires disciplined service in return. The resulting *mansabdari* system tied each *sipahi's* social position to precise obligations. As one modern account notes, "the effective striking arm of Mughal power lay in its cavalry," and therefore each *mansabdar* was obliged to "maintain well-equipped cavalry contingents with approved breeds of horses[9]." The state paid these officers largely in *jagir* (land revenue) revenue instead of cash. Since both rank and pay could be granted or withdrawn by the emperor, *mansabdars* remained firmly dependent on central favor. In short, Mughal *sipahi* identity was built into a bureaucratic framework of rank, land, and patronage from the start, a sharp contrast to the samurai's feudal heritage.

Codification and the Formation of Both States

The two states codified warrior norms in very different ways. After unification in Tokugawa Japan, the shogunate needed to control a peace‑time samurai class. The *Buke Shohatto* (1615) became the cornerstone of this effort. Its articles are social rather than strictly military. It even limits private castle building and marriage alliances: "*Marriage must not be contracted in private…To form a factional alliance through marriage is the root of treason*." Lords and retainers were bound by strict etiquette to the point where even their dress was regulated. Notably, Article 12 commands that "*the samurai of all domains must practice frugality…[for] there is nothing which will corrupt public morality more than this*[10]." This edict and others subordinated personal ambition to public order, replacing the old autonomy of military houses with bureaucratic predictability. In practice, castles, castles roads, and inspections (e.g. *sankin-kōtai*) enforced compliance. Later revisions (in 1635) tightened these rules further. These regulations worked in conjunction with *sankin-kōtai*, the alternate attendance system that required daimyo to spend regular periods at Edo while maintaining costly residences there[11]. Although primarily directed at domain lords, the system indirectly disciplined samurai retainers by tying their fortunes to the political supervision of their households.

Importantly, actual soldiering was de-emphasized. The injunction to study martial arts appears in the code, but largely as a reminder of status. The real focus was on decorum and duty. To cement this shift, didactic works like Yamamoto Tsunetomo's *Hagakure* (early 18th century) moralized loyalty as a personal virtue. In it, a retainer is taught to serve his master "without calculation, as though already dead[12]," abandoning any concern for reward or protection. Such texts were nothing but pedagogical devices that naturalized the Tokugawa order by instilling unconditional loyalty into warrior consciousness.

---

8 Abu'l Fazl ibn Mubarak. *The Akbarnama*. Translated by Henry Beveridge, Asiatic Society, 1907–1939. 3 vols. Pp 1:14–18.

9 Richards, John F. *The Mughal Empire*. Cambridge UP, 1993, pp 67–75.

10 "Buke Shohatto," p. 395.

11 Vaporis, Constantine Nomikos. *Tour of Duty: Samurai, Military Service in Edo, and the Culture of Early Modern Japan*. U of Hawai'i P, 2008.

12 Yamamoto, Tsunetomo. *Hagakure: The Book of the Samurai*. Translated by William Scott Wilson, Kodansha International, 1979, pp 17–21.

Contrariwise, Akbar's Mughal state formalized warrior service through fiscal administration. The *Ain-i-Akbari* (1590s) devotes chapters to the *mansabdars*—it fixes pay scales, *sawar* (cavalry quotas) , and household obligations for each rank. To enforce this, Akbar instituted *dagh* (horse-branding) and *chehra* (descriptive rolls). Military duty became an auditable function of the empire. Rather than preach loyalty abstractly, the Mughal regime enforced it through incentives and monitoring. Each *mansabdar's jagir* and salary depended on meeting these numeric standards. If a noble failed muster or displeased the court, the emperor could reduce his rank or reassign his *jagir*[13]. Through these means, martial obligations were integrated into the revenue system. In sum, Tokugawa codification relied on moral pedagogy and personal hierarchy, whereas Mughal codification translated warrior norms into bureaucratic accountability embedded in the fiscal-military order.

Loyalty and the Structure of Service

The concept of loyalty worked differently in each system. In Japan, loyalty was framed as an ethical imperative of the household. Samurai were expected to serve their lord without question or self-interest. *Hagakure* exemplifies this absolutism. It teaches that a retainer should be ready to die at any moment for his master. Such rhetoric intentionally overlooks historical reality. In the Sengoku age, vassals often switched sides when a lord failed them or when personal safety demanded. Under Tokugawa rule, however, no excuse for disloyalty was tolerated. Samurai loyalty became an inner discipline, expressed through etiquette and obedience rather than by the mutual contracts of war. Indeed, the fervency of Tokugawa moralism suggests underlying anxiety as most samurai of the time subsisted on fixed stipends, many were in debt, and opportunities for advancement were few. By elevating loyalty into virtue, the regime compensated warriors for their restricted agency and silenced potential grievances.

However, in the southern empire of Hindustan, loyalty was organized structurally. Each *mansabdar* knew that his rank and *jagir* depended on imperial favor, so allegiance was always conditional. The transferability of jagirs was particularly significant. By regularly reassigning revenue holdings, the Mughal state limited the ability of military elites to establish durable territorial bases independent of imperial authority. Because the *mansabdari* system allowed promotion, demotion, and transfer, nobles had incentive to negotiate. During succession crises (e.g. the wars after Shah Jahan's illness, 1657-58), many officers delayed support until the outcome was clear, effectively bargaining by pledging allegiance to the winner[14]. This was not seen as dishonorable but as political calculation within the system. The Mughal solution was bureaucratic. Victorious emperors redistributed ranks and *jagirs* to reward those who had been loyal or opportunistic on the winning side. In this way, loyalty was restored through administration. Unlike Tokugawa Japan, where disloyalty was cast as a moral failing, the Mughal state simply recalibrated the *mansab* hierarchy. In both cases, the persistence of self-interest is apparent. Warrior ethics could guide behavior, but could not eliminate power politics.

---

[13] Richards, John F. *The Mughal Empire*. Cambridge UP, 1993.

[14] Alam, Muzaffar, and Sanjay Subrahmanyam, editors. *The Mughal State*, 1526–1750. Oxford UP, 1998.

Military Order and Technology

Practical warfare shaped how discipline was enforced. Mughal armies of Akbar's reign combined large cavalry contingents with emerging gunpowder weapons. Campaign reports in the *Akbarnama* stress logistics and order. For example, during Akbar's Gujarat expedition (1572-73) Abul-Fazl notes the emperor's concern with supplies, troop dispositions and chain of command more than lone heroism[15]. The state's technical demands reinforced bureaucratic discipline as a *mansabdar* had to account for every horse and trooper, and cannons and matchlock infantry were kept in strict muster. Commanders used *shikargah* (gazetteers) and *chehra* listings to ensure equipment quality. In short, Mughal discipline was external, or verifiable by the records, fitting an army that relied on coordinated arms and imperial resources.

Japan's situation was reversed. By the early seventeenth century, matchlocks had become common, and the decisive wars (Sekigahara, Osaka) confirmed the shogunate's power. With two centuries of peace ahead, the samurai's fighting role diminished. Standing patrols and castle guards replaced field armies. As a result, Tokugawa discipline shifted from martial training to conduct. The *Buke Shohatto* devoted little space to actual drill. Even when urging martial study, it functioned more as a status reminder. For example, Article 1's admonition to cultivate military arts comes amid a string of injunctions about order and frugality. By the mid-1600s, etiquette and paperwork took the place of swordsmanship as proof of a warrior's worth. Samurai were evaluated on their performance of ceremonies, archery displays, legal knowledge, and obedience to protocol. In effect, the regime neutralized the samurai's combat value by channeling discipline into bureaucratic and ritual norms.

Ideology versus Practice

Despite the idealism of their rhetoric, both systems faced practical limits. Moral absolutism in samurai lore masked the shogunate's real constraints. The *Hagakure* demand that a warrior live "already dead" erases any allowance for material needs. Yet actual samurai often found such devotion hard to sustain. Throughout the Sengoku era, retainers defected or banded together when lords failed to provide land, stipends, or security. Contemporary documents, from service oaths to rice grant charters, treat loyalty as negotiable. By silencing this history, Tokugawa ideology precluded even the language of contractual obligation. In this tension lies an irony: the more the samurai were told to act like moral saints, the more it signaled the regime's need to compensate for unresolved grievances. Zen-like exhortations took the place of honest bargaining.

The Mughal case shows a parallel gap. Political theory (e.g. the doctrine of *ṣulḥ-i kull* or universal peace, a secular religious policy of Akbar's) emphasized obedience, justice, and universal order, but nobles routinely acted in self-interest. For instance, during succession struggles many *mansabdars* shifted allegiance opportunistically[16]. This maneuvering was not

[15] Abu'l Fazl ibn Mubarak. *The Akbarnama*. Translated by Henry Beveridge, Asiatic Society, 1907–1939. 3 vols, pp 2:102–110.

[16] Chandra, Satish. *Medieval India, Part II*. Har-Anand, 2005, pp 247–254.

corruption so much as the system's flexibility. Victorious rulers recalibrated the state instead of spreading propaganda, simply reshuffling ranks and *jagirs* to absorb or discard opponents. Thus, loyalty was "enforced" by material measures rather than moral suasion. In both empires the persistence of elite self-interest underlines the point that warrior ethics could direct behavior but never fully override personal calculation.

Comparative implications for Governance

In both Tokugawa Japan and Mughal India, ruling elites used the language of the warrior to manage potentially dangerous armies, but the mechanisms differed with each state's political economy. In Tokugawa Japan, a long peace rendered mass mobilization unnecessary while preserving a large samurai class. The shogunate's solution was to keep the samurai in place as an upper stratum but strip away their military autonomy. Through edicts like the *Buke Shohatto* and manuals like *Hagakure*, martial virtue was re-framed as moral discipline and loyalty within a settled hierarchy. Samurai status persisted, but its content became obedience, frugality, and ritual propriety. Crucially, this discipline was internalized as daimyo families became responsible for policing their own retainers. The shogunate thus achieved stability without vast standing armies or expenditure.

The Mughal Empire took a different path suited to its continuous wars and expansion. Rather than teaching nobles to lay down their arms, Akbar integrated them into a fiscal-military framework. The *mansabdari* system bound loyalty to rank, pay, and imperial service. Procedures such as horse-branding and muster inspections translated martial promise into measurable compliance. In effect, authority rested not on transforming a warrior's inner character but on the emperor's ability to allocate *jagirs*, promotions and appointments. The system welded civil administration and military command into one framework. In practice, Japanese samurai learned to prove themselves by self-restraint, whereas Mughal *sipahis* proved themselves by meeting administrative quotas.

The persistence of later *Bushidō* narratives has often obscured the historical specificity of Tokugawa warrior ethics. Meiji-era writers such as Nitobe Inazō, by retrospectively transforming Tokugawa notions of service and discipline into a timeless national ethic, masked the extent to which these ideals emerged from a project of pacification to suit the militaristic needs of the budding Japanese empire.

In sum, the warrior ethics of Tokugawa Japan and the Mughal Empire were not timeless moral codes but historically contingent strategies of rule. Although both samurai discourses and Mughal rhetoric praised loyalty, discipline and duty, each ideology was molded by specific administrative needs. Tokugawa authorities turned martial language into moral training and household regulation, using ethical discipline to neutralize an armed class in peacetime. In contrast, the Mughal state absorbed military elites into a graded system of rank and inspection, tying loyalty to fiscal accountability and battlefield readiness. By reading prescriptive texts like *Buke Shohatto*, *Hagakure*, and the *Ain-i-Akbari* alongside campaign records, we see how ideals of service were operationalized through very different institutional technologies. In the end, both

cases demonstrate how martial identities were manufactured to legitimate power and control force.

Work Cited

Abu'l Fazl ibn Mubarak. The Akbarnama. Translated by Henry Beveridge, Asiatic Society, 1907–1939. 3 vols.

Abu'l Fazl ibn Mubarak. The Ain-i-Akbari. Translated by H. Blochmann and H. S. Jarrett, Asiatic Society of Bengal, 1873–1907. 3 vols.

Alam, Muzaffar, and Sanjay Subrahmanyam, editors. The Mughal State, 1526–1750. Oxford UP, 1998.

Babur. The Baburnama. Translated by Wheeler M. Thackston, Modern Library, 2002.

"Buke Shohatto" ["Laws for the Military Houses"]. Sources of Japanese Tradition, edited by Wm. Theodore de Bary et al., vol. 1, Columbia UP, 2001.

Chandra, Satish. Medieval India, Part II. Har-Anand, 2005.

Richards, John F. The Mughal Empire. Cambridge UP, 1993.

Turnbull, Stephen. Samurai Warfare. Cassell, 1996.

Vaporis, Constantine Nomikos. Breaking Barriers: Travel and the State in Early Modern Japan. Harvard University Asia Center, 1994.

Vaporis, Constantine Nomikos. Tour of Duty: Samurai, Military Service in Edo, and the Culture of Early Modern Japan. U of Hawai'i P, 2008.

Yamaga, Sokō. The Way of the Samurai. Translated by Thomas Cleary, Shambhala, 2007.

Yamamoto, Tsunetomo. Hagakure: The Book of the Samurai. Translated by William Scott Wilson, Kodansha International, 1979.

Research Analysis on the Relationship Between Formal and Informal Education and the Income of Youth from SHE Clan Families in Tianchang City By Garrett Fulin Zhao

Abstract

Based on survey data from 55 households of young members of the SHE clan families in Tianchang City of China, this study examines the relationships between educational attainment and income levels, educational attainment and comprehensive social security benefits, as well as informal education and income. The findings indicate that individuals with a bachelor's degree or higher account for a higher proportion (68.4%) of those in the high-income bracket compared to those with high school education or below (30.0%), demonstrating a positive correlation between educational attainment and income; among bachelor's degree holders, the proportion enjoying basic social security benefits or higher is higher (84.2%) than among those with lower education levels (45.0%); additionally, participants who have engaged in informal education three or more times represent a greater share of the high-income group (61.1%) compared to non-participants (28.6%). However, the study data exhibit selective reporting bias, and conclusions should be drawn with caution.

Keywords Educational attainment; Income; benefits; Informal education

1. Introduction

Whether education can effectively improve individual employment quality and income levels has long been a key issue of concern in the fields of educational economics and sociology. In modern society, academic education is widely regarded as a crucial means of enhancing human capital and boosting individual competitiveness in the labor market. However, in reality, there exists a phenomenon where some individuals with lower educational backgrounds achieve higher incomes, while others with higher education face employment challenges. This has contributed to the persistence of the notion that "education is useless" among the public and has blurred their understanding of education's social functions.

In response to the aforementioned controversies, the academic community has conducted extensive empirical research on the relationship between education and income. The wage determination equation proposed by Mincer (Mincer, 1974) provides a classic theoretical framework for the quantitative analysis of educational returns. Existing literature consistently demonstrates a significant positive correlation between educational attainment and individual income (Fang,and Ding, 37), based on data from the China Family Panel Study, found that educational attainment not only directly enhances individual wage income but also indirectly promotes wage growth by improving cognitive abilities such as computational skills, with the mediating effect of computational ability reaching as high as 63.4%. Yang, Jinfeng and Zhang, Qingxia (94) using survey data from rural households in Shanxi Province, discovered that for every additional year of education among migrant workers, their average income increases by 5.1%, and receiving technical training further contributes to higher income levels. Hou, Haibo

and Liu,Yahui (40) further noted that vocational education has a more pronounced impact on income improvement for informal workers. These studies collectively indicate that educational attainment plays a multifaceted role in income formation.

However, existing research predominantly relies on data from large-scale national or regional sampling surveys, with relatively limited attention paid to the relationship between education and income within specific communities. This study conducts a confirmatory analysis of the correlation between educational attainment and income levels using resident education and income data collected from a survey of young members of the SHE clan families in Tianchang City of China. Additionally, the survey incorporates the variable of informal education —specifically, the frequency of training or learning activities undertaken by respondents after entering the workforce—to examine how different types of learning experiences influence income formation. It should be noted that the data were collected through oral interviews with elderly participants, who exhibited a selective bias when reporting their children's educational backgrounds and income status, particularly by being more inclined to mention children with better educational outcomes; this limitation will be further elaborated later.

The subsequent structure of this paper is as follows: Part II presents the data sources and sample description; Part III outlines the research findings, including descriptive statistics and cross-analyses; Part IV&V comprises discussions and conclusions.

2. Data Sources and Sample Description

The survey data for this study were obtained through door-to-door interviews conducted with senior members of the SHE clan families in Tianchang City of China, who provided oral accounts regarding their children's educational background, employment status, and income levels. A total of 56 questionnaires were collected; after excluding one invalid sample (due to incomplete information), 55 valid samples were included in the analysis. The variables covered seven dimensions: age, education level, employment status, field of specialization, frequency of participation in informal education programs, monthly income range, and overall financial security level. It should be noted that due to the use of non-random sampling and respondents' tendency to report cases where their children achieved better outcomes, the data exhibit a degree of selection bias—high-income and highly educated samples may be over-represented. Therefore, the conclusions of this study should be interpreted as observations of trends within the valid sample population rather than precise generalizations about the overall situation.

3. Findings of the Research

3.1 Basic Sample Characteristics

Among the 55 valid samples, age distribution showed a predominance of individuals aged 36–45 (27,49.1%), followed by those aged 29–35 (14,25.45%). Educational background comprised 27 respondents (49.1%) with high school education or below, 22 (40.0%) holding bachelor's degrees or higher, and 6 (10.9%) with associate degrees. Employment status included

41 full-time employees (74.5%), 9 students (16.4%), and 3 freelancers or self-employed individuals (5.45%). Overall, the sample primarily consisted of middle-aged individuals with high school education or below who were employed full-time.

3.2 The Relationship Between Educational Background and Income

The following statistics cover full-time employees, freelancers, and self-employed individuals (excluding students), totaling 44 people.

Education Level	3000-5000	5000-8000	8000-12000	12000+	Subtotal
High school and below	2	4	5	12	23
Specialty	0	2	1	3	6
Undergraduate Program	0	0	4	6	10
Key Undergraduate Prograr / Double First-Class Initiati	0	0	2	3	5
Total	2	6	12	24	44

Regarding income distribution, among individuals with a high school education or lower, 8.7% (2/23) earn less than 5,000 yuan; none of those with a bachelor's degree or higher earn below this level. In the income bracket below 8,000 yuan, high school graduates or less account for 26.1% (6/23), whereas those with a bachelor's degree or higher represent 0% (0/15). Within the high-income range above 12,000 yuan, the proportion of individuals with a bachelor's degree or higher is 60.0%, higher than that of those with a high school education or lower (52.2%). These findings indicate a positive correlation between educational attainment and income within this sample population.

This study also aims to provide a reference for decision-making regarding the establishment of a youth scholarship and financial aid program by the SHE Clan Association in Tianchang City.

3.3 The Relationship Between Academic Qualifications and Comprehensive Support

The following statistics cover the employed population (excluding students), totaling 44 individuals.

Education Level	Guarantee	Guarantee	Guarantee	Guarantee	Subtota
High school and below	12	3	10	2	27
Specialty	0	1	5	0	6
Undergraduate Prograr	4	1	2	7	14
Key Undergraduate Programs / Double First-Class Initiatives	2	0	3	2	7
Total	18	5	20	11	44

In terms of social security coverage levels, among individuals with high school education or below, 55.6% (15/27) receive no benefit or only salary benefits, whereas among those with a bachelor's degree or higher, 66.7% (14/21) have coverage level 2 or above (i.e., at least basic social security benefits). Data show that the proportion of individuals without any social security benefits (level 0) among those with high school education or below is significantly higher (44.4%) than among those with a bachelor's degree or higher (28.6%), indicating that lower-educated groups face relatively greater challenges in employment security.

3.4 The Relationship Between Informal Education and Income

Participation in informa education	3000-5000	5000-8000	8000-12000	12000+	Subtotal
0 times	1	0	0	1	2
1–2 times	1	5	5	12	23
3 times or more	0	1	7	11	19
Total	2	6	12	24	44

Among the 44 employed respondents, 43.2% (19/44) had participated in informal education three or more times, 52.3% (23/44) participated once or twice, and only 4.5% (2/44) did not participate at all. Within the high-income group of over ¥12,000,57.9% (11/19) had participated three or more times, while 52.2% (12/23) participated once or twice. Due to the small sample size for non-participants (only two individuals), a meaningful comparison of their income distribution is not feasible. Overall, there may be a positive correlation between participation in informal education and higher income; however, the discriminative power of this variable in this sample is limited, and the findings should be interpreted with caution.

Discussion

4.1 Main Findings and Literature Comparison

The findings of this study are largely consistent with existing research conclusions. The study by Fang, Chao and Ding,Jie (37) demonstrates that formal education indirectly promotes wage growth by enhancing cognitive abilities, and our investigation also observed a similar positive correlation between education level and income within a narrower scope. Jinfeng Yang and Qingxia Zhang (94) found that each additional year of education increases migrant earnings by 5.1%; similarly, in this study, the proportion of individuals with a bachelor's degree or higher in the high-income bracket was 60.0%, significantly higher than that of those with high school education or below (52.2%), reflecting a consistent trend. Furthermore, our study revealed that individuals who participated more frequently in informal education had a higher share in the high-income group, aligning with Hou Haibo and Yahui Liu 's (40) conclusion that vocational education and training have a positive impact on income.

4.2 Discussion on Research Limitations and Bias

The data collection for this study employed oral interviews conducted under convenience sampling: first, the respondents were elderly members of the SHE family youth in Tianchang City who were readily accessible to the researchers; second, the information provided by the respondents concerned details about their children. These two sampling and reporting methods collectively constitute the primary limitations of this study.

Regarding sample representativeness, convenience sampling means that the study's sample was not drawn randomly from the target population but was limited to the researcher's accessible scope. This implies that the sample distribution across dimensions such as education level, income, and geographic location may not accurately reflect the overall population profile. In practice, respondents tend to report cases of children with successful outcomes more frequently (e.g., higher education attainment, stable income, employment in major cities), while rarely mentioning children with average or poor outcomes. This phenomenon of "selective reporting" is common in oral surveys, reflecting a psychological tendency among respondents to maintain a positive family image in social interactions.

Regarding income structure, it is essential to categorize income by specific categories such as wage income and business income. The lack of detailed classification has resulted in insufficient new insights from the survey.

Regarding the specific manifestations of bias, from a sample structure perspective, individuals with a bachelor's degree or higher account for approximately 40% of this survey participants, while those in the high-income bracket (over ¥12,000) make up about 55%, both significantly higher than the national average. This distribution pattern indicates that selection bias has indeed influenced the composition of the sample—specifically, high-educated and high-income groups are over represented. Consequently, the positive correlation between education level and income observed in this study may be somewhat overstated.

It should be noted that the presence of selection bias does not negate the fundamental conclusion that there is a positive correlation between educational attainment and income; rather, it reminds readers to exercise caution when interpreting specific numerical values. The findings of this study should be understood as a "collective observation of positive cases" rather than an accurate inference about the overall state of the labor market.

In future research, it is advisable to combine random sampling with cross-validation using multi-source data (e.g., official income statistics and corporate personnel records) to further examine the underlying mechanisms linking education and income, thereby reducing the bias risks associated with oral surveys.

5. Conclusion and Recommendations

5.1 Conclusion

Based on the above analysis, this study draws the following conclusions:

First, educational attainment has a positive impact on income. Within this sample, the

proportion of individuals with a bachelor's degree or higher in the high-income bracket (60.0%) is higher than that of those with a high school education or lower (52.2%), indicating that higher education levels are associated with greater likelihood of entering the high-income range. Additionally, there is a clear positive correlation between educational attainment and employment security: among those with a bachelor's degree or higher, the proportion enjoying basic social security benefits or higher (66.7%) exceeds that of those with a high school education or lower (44.4); conversely, the proportion without any social security benefits among those with a high school education or lower (44.4%) is significantly higher than that among those with a bachelor's degree or higher (28.6%). This demonstrates that educational attainment not only influences income levels but also affects employment quality.

Second, there is a positive correlation between informal education (such as vocational training and on-the-job learning) and higher income. Those who participated in three or more forms of informal education account for a slightly larger proportion of high-income earners compared to those who participated once or twice, suggesting that continuous learning may have a positive impact on income. However, due to the small sample size of participants who did not engage in informal education (only two individuals), the robustness of this conclusion is limited.

Third, the study's sample exhibits significant selective reporting bias. Due to the survey methodology (covering a broad elderly population via convenience sampling), individuals with high incomes and advanced education may be over represented; thus, the findings should be interpreted as "convergent observations of positive cases" rather than precise estimates of overall labor market conditions. Future research should employ more rigorous sampling designs and multi-source data validation to further examine the underlying mechanisms linking education and income.

5.2 Recommendations

This study reveals that individuals with lower educational backgrounds are at a relative disadvantage in terms of income levels and employment security. Among those with high school education or below, 55.6% lack job security or receive only fixed salaries, compared to 33.3% among those with bachelor's degrees or higher; within the group without job security, the proportion of low-educated individuals is significantly higher than that of highly educated ones. This phenomenon is particularly pronounced among flexible workers and those engaged in informal employment. Therefore, the government should increase targeted investment in vocational training and continuing education for low-educated groups, especially in industries such as manufacturing and services that employ large numbers of them. Additionally, it should explore establishing simplified social insurance enrollment procedures for flexible workers, lowering contribution thresholds, enhancing social security coverage for low-educated groups, and narrowing the employment quality gap between different educational backgrounds.

It is recommended that the Tianchang SHE clan Association establish scholarships and financial aid programs to encourage more young members of the clan to pursue both formal and non-formal education, while providing small grants to students from extremely disadvantaged

families to enable them to complete their formal education. Through sustained efforts over generations, education can significantly improve the economic conditions of the Tianchang SHE clan families.

The sample for this study was drawn using convenience sampling, and oral interviews are subject to selective reporting bias, with highly educated and high-income individuals being over represented, which may lead to an overestimation of the positive correlation between education level and income. Future research could employ random sampling methods combined with cross-validation using official statistical data (e.g., statistical yearbooks, social security records) to mitigate bias risks.

Additionally, this study did not collect information on parents 'educational backgrounds, making it impossible to analyze intergenerational educational transmission effects. Existing research indicates that parental education levels significantly influence children's educational opportunities and income levels. Future studies should incorporate parental educational background into their analytical framework to more comprehensively elucidate the mechanisms linking education and income.

Works Cited

Fang, Chao, and Ding, Jie. From Human Capital to National Knowledge Capital: The Impact of Academic Education and Cognitive Ability on Labor Wage Income. Chongqing Higher Education Research, 2025.

Hou, Haibo, and Liu,Yahui. Heterogeneity analysis of educational returns in informal employment. Survey World, 2018.

Mincer, J. . Schooling, Experience, and Earnings. New York: Columbia University Press, 1974.

Yang, Jinfeng, and Zhang, Qingxia. The impact of education on rural laborers' employment and income: A survey-based study in Shanxi Province. Higher Agricultural Education, 2005.

From Sequence to Function: Programmable Structural Proteins as Sustainable Alternatives to Conventional Plastics By Lian Huan

Abstract

Conventional plastics are widely used due to their strength, durability, and processability; however, the limitations of current disposal methods call for more sustainable alternatives. This paper examines structural proteins as potential alternatives for petroleum-based plastics by comparing glycinin 11s with polyethylene (PE), suckerin-12 with polyethylene terephthalate (PET), and major ampullate spidroin 1 with nylon 6,6. These comparisons underscore the idea that through amino acid sequences and molecular structure, protein-based materials are able to achieve plastic-like qualities. Furthermore, this sequence-based relationship allows modifications on a molecular level, giving them a more precise form of tunability compared to plastics. Nonetheless, these materials are still facing challenges that prevent them from being manufactured on an industrial level: production efficiency, cost, and environmental stability are still major limitations.

Introduction

The study of proteins spans from examining the most basic structures to a more complex understanding of their role in biological life. Many early breakthroughs on the structure of proteins, like Frederick Sanger's success in sequencing the protein insulin and Christian Anfinsen's experiments on the protein ribonuclease, determined a governing feature of all proteins: structure determines function (Berg; Kresge et al.). As technology has advanced and become more accessible, the Human Genome Project revolutionized the study of proteins, providing the genetic basis of thousands of proteins in humans (Collins and Fink). Moreover, advanced techniques like cryo-electron microscopy and AI-based protein folding models (e.g., AlphaFold 2) are now revealing detailed protein structures directly from sequence, dramatically accelerating research. With these powerful tools, a deeper understanding of protein form and function has been elucidated and expanded beyond biology, like in the search for replacements for synthetic polymers or potential industrial reactants. Notably, one of the most promising applications for these macromolecules is their potential as an alternative to plastics.

Plastics are synthetic polymers constructed from chains of repeating monomers. These chains may vary in length: longer chains provide more tensile strength and a higher melting point, while shorter chains are more flexible with a lower melting point (Yang et al.). In addition, how orderly packed these chains or their crystallinity are also factors that influence their functions and properties. More crystalline polymer structures manifest as more rigid plastics, while less crystalline structures lead to more flexible plastics. In some cases, these chains also exhibit branching, where monomers begin to branch off the main polymer backbone. Branching can effectively reduce the crystallinity of the overall structure, making it more amorphous and flexible (Chremos and Douglas). Lastly, cross-linkage may also occur between neighboring chains. These linkages can be facilitated by covalent bonds or weaker hydrogen bonds; however,

in both cases, these interactions create a chemical network that improves the durability and gives a more rigid structure to the plastic (Chen et al.).

Proteins are constituted by repeating monomers known as amino acids (AA), with each of the existing 20 AAs characterized by a specific R group, which is a specific side chain that determines the AA's properties (Stollar and Smith). By joining these amino acids together, a polypeptide chain is formed, and this is the primary structure of a protein. This chain can then undergo conformational changes that are facilitated by hydrogen bonds. These structures are known as secondary structures and are characterized by two major types of structures formed: alpha-helices or beta-pleated sheets. The tertiary structure is characterized by the interactions of specific AA's R groups and interactions of nonpolar AAs. As there may be hydrophobic AAs that exist in the chain, and in many cases these proteins are working in an aqueous solution, these sections will start to aggregate towards the center of the protein, while the hydrophilic AAs begin to move outwards, to surround the protein (Stollar and Smith). Finally, the quaternary structure of a protein is where multiple strands of polypeptide chains join together to form a larger complex. This complex is joined non-covalently through hydrogen bonds, ionic bonds, or Van der Waals forces.

The dominance of synthetic polymers, especially plastics, in modern manufacturing comes at a devastating and escalating environmental cost. Their resistance to degradation leads to centuries-long accumulation in ecosystems, with millions of metric tons of them entering oceanic and terrestrial environments annually (Lin et al.). The chemicals associated with plastics, such as lead, cadmium, and phthalates, can have catastrophic effects if exposed to or ingested by living organisms (Lin et al.). Phthalates, for example, a series of chemicals used as a plasticizer, can cause severe damage to the endocrine system, inflicting long-term negative impacts on reproduction and pregnancy (Wang and Qian). The biological nature of proteins alleviates the environmental concerns brought up by plastics. Unlike plastics derived from petroleum or other chemicals, proteins are composed of naturally occurring amino acids that are renewable. Furthermore, the function of these proteins is not a result of chemical additives, but is the result of the interaction between the sequence of the amino acids.

Therefore, structural proteins have become the center of focus for becoming the future substitutes for plastics. However, while these functional properties are essential for creating viable alternatives, the molecular principles governing how specific sequences dictate them remain poorly understood. How specific sequences of amino acids in structural proteins create the unique functional properties (e.g., biodegradability, tunable strength, self-assembly) observed in biological alternatives to conventional plastics is vital to our understanding and future research for alternatives. By comparing and contrasting three types of plastics and their potential protein replacements, this paper provides an overview of the relationship between amino acid sequences and plastics.

Glycinin 11S as a Biodegradable Answer to Polyethylene Packaging

Glycinin, a major type of soybean protein, may be a functional replacement for polyethylene (PE). Glycinin 11S, along with β-conglycinin, is a major storage protein for soybeans, especially for germinating seeds, before they can perform photosynthesis (Singh et al.). The protein does this by storing nitrogen and amino acids that are critical for the growth of these plants. The rich content of glutamine in the primary structure of the protein allows it to hold onto nitrogen atoms due to the amide groups on the amino acid (Sharma and Sharma). Additionally, the polypeptide chain has a higher proportion (46.7%) of hydrophilic AAs, such as glutamine and asparagine, than hydrophobic AAs (23.5%) such as valine, leucine, and alanine (Peng et al.). The secondary structure of glycinin 11s is facilitated through hydrogen bonds (Peng et al.). The prevalence of these bonds gives glycinin 11s a structure that is dominated by beta-structures and random coils with 5.2% of α-helices, 34.8% β-structures, and 60% random coils (Peng et al.). These secondary structures are then folded and packed to form the tertiary structure, where they fold into a compact globule with a hydrophobic core and hydrophilic surface (Adachi et al.). The core structure of this protein is a hexamer and is composed of different combinations of 5 types of isoform subunits in soybeans (A1aB1b, A1bB2, A2B1a, A3B4, and A5A4B3) (Maruyama et al.). These subunits are each individually formed by two chains (12 polypeptide chains in total): acidic A chains and basic B chains. These two chains are bonded together via disulfide bonds between cysteine residues, and this stabilizes the protein in its native and mature state (Singh et al.).

On the other hand, polyethylene is a type of synthetic resin that is widely used in the food industry for food wraps and shopping bags, and in the automobile industry for fuel tanks (H. Zhang). As the name suggests, polyethylene is manufactured through the polymerization of ethylene monomers $(–CH_2–CH_2–)n$. This is achieved through three processes: initiation, propagation, and termination. The process begins with initiation. After applying heat, the organic peroxide is separated into two free radicals (Kehinde et al.). These radicals then begin to react with ethylene. One radical binds to an ethylene monomer, which breaks the previous double bond between two ethylene monomers to form single bonds (Kehinde et al.; dos Santos). This propagation action continues as more and more ethylene molecules are bonded to the new radical via single bonds. This polymerization process stops when the last radical on one chain bonds to a free radical on another chain, which forms the final product – polyethylene (Kehinde et al.; dos Santos). On this basis, polyethylene may also contain branches that branch off the main polymer backbone, and are known as low-density polyethylene (LDPE). Conversely, those that don't branch off, and are linear, are called high-density polyethylene (HDPE) (dos Santos). The branches in LDPE are formed at high pressure (usually around 150–350 MPa) and a temperature of 80–300 °C, while linear HDPE is formed at relatively lower pressure (Pham; Emblem). The differences in structures lead to differences in function and properties in both kinds of PE. LDPE, due to its branched structure, cannot be packed very tightly together to form crystalline structures, granting the material high flexibility (dos Santos). On the contrary, HDPE's linear structure allows for easy packaging into orderly, crystalline structures, resulting in stronger tensile strength and moderate stiffness.

Glycinin 11s is a potential alternative for PE due to the comparable properties that the key residues of glycinin 11s possess. Despite glycinin containing a relatively higher proportion of hydrophilic amino acids, the hydrophobic core of the protein establishes itself as an "internal maze" (Cho and Rhee). The high content of glutamine in this protein leads to a rich amount of β-sheets and turns. These β-structures lead to a reduction in oxygen permeability due to the reduced amount of volume that allows water molecules to pass. This is similar to PE; being a hydrocarbon, it is entirely non-polar, which naturally resists the passage of water molecules (Klute and Franklin). Furthermore, the β-sheets formed by the hydrophilic AAs like glutamine and asparagine also contribute to the mechanical strength of the overall protein (Perutz et al.; Cohen and Eisenbach). Research shows that the higher the β-sheet content, the denser the network structure of the protein, which leads to a higher stiffness and mechanical strength (Cohen and Eisenbach). Soy-protein-based films have been studied for mechanical strength and water-vapor permeability (Cho and Rhee). Similarly, PE's mechanical strength is influenced by chain structure, crystallinity, processing technique, and temperature; HDPE can reach around 20–40 MPa depending on processing conditions (Shi et al.; Amjadi and Fatemi). Although the mechanical strength of PE is slightly higher than that of glycinin 11s, it is recognized that proteins are tunable in structure, allowing for the introduction of mutations to change the protein structurally to fit a better purpose (Pey et al.). Unlike PE, glycinin allows for higher degrees of tunability in structure (Hu et al.). For example, through four industrial modification methods: heat treatment (H), alkaline treatment (A), glycosylation (G), and oxidation (O), which were applied to glycinin, there was a significant change in the secondary structures of the protein (Hu et al.). Notably, there was a decrease in the content of α-helices, while an increase in the amount of β-sheets and turns. This is mostly attributed to the denaturing and unfolding process of the protein under these conditions, which led to the formation of new, stable β-structures (Hu et al.). These new secondary structures are then associated with improved mechanical properties, making glycinin a suitable alternative to PE.

Architectural Comparison between Suckerin-12 and Polyethylene Terephthalate

To further demonstrate proteins' viability as an alternative for plastics, suckerin-12 is compared with polyethylene terephthalate. Squid ring teeth are found along the predatory tentacles in Decapodiform cephalopods, which include squid and cuttlefish, around the interior of their suckers that allow for grip strength for predation (Hiew and Miserez). The proteins that construct these appendages are called Squid Ring Teeth (SRT) proteins, and they belong to a family of proteins called suckerins (Hiew and Miserez). Suckerins are a block co-polymer protein, composed of two or more different polymer chains linked together, that self-assemble into the supramolecular network that forms SRT (Hiew and Miserez). Nearly all proteins from this family contain repeating sequence motifs that resemble [Pro-M1-Pro-M2], resulting in alternating crystalline (M1) and amorphous regions (M2) in the protein structure (Sun and Ding). Suckerin-12 (S12) is one of the most thoroughly studied proteins in this family and possesses remarkable material properties. The majority of the sequence of S12 in M2 is GGX motifs (X=

any AA), where X is most commonly associated with tyrosine (Y) (Liu et al.; Hershewe et al.). Beyond the GGX motifs, the M1 module of S12 contains a rich amount of histidines (H), which contribute to the pH dependence of S12 (Hershewe et al.). Histidine's side chain is an imidazole group, which has a pKa of around 6.3, relatively close to neutral pH (Calinsky and Levy). This facilitates histidines to switch between neutral and positively charged states based on the pH environment around it, directly impacting structures in S12 (Calinsky and Levy). Due to the prevalence of glycine in the GGX motifs, this allows for blocks of amorphous regions and random coils to form, as the R group of glycine is just a single hydrogen. This allows for the chain to avoid locking up into rigid secondary structures like α-helices or β-sheets (Liu et al.). Unlike glycine with its simple R group, tyrosine's R group is more complex, where a hydroxyl group is attached to an aromatic benzene ring (phenolic hydroxyl group) (Hershewe et al.). The high content of tyrosine and its phenolic hydroxyl group allows for the supramolecular arrangement of S12 through π···π stacking that assists in stabilizing remote amorphous regions of the protein. Additionally, these interactions also form interprotein di-Tyrosine cross-linkages with β-sheets to allow for covalent connections with separate β-sheet-rich monomers, enhancing stability and structural support (Hershewe et al.). Individually, S12 is constructed by six β-sheet structures rich in alanine (A) and histidine (H) residues stabilized by hydrogen bonding, and is flanked by amorphous blocks (Hershewe et al.). The β-sheets are separated by proline residues, as they are able to serve as capping mechanisms that prevent elongation of these structures (Shamsir and Dalby). Thus, the protein can self-assemble structures that alternate between nano-scale crystalline structures and amorphous, flexible regions, resulting in high mechanical strength and flexibility.

After examining how S12 structurally achieves functional toughness and flexibility, it is insightful to see how synthetic materials accomplish the same purpose, of which the most commonly seen is polyethylene terephthalate (PET). PET is a type of aliphatic-aromatic thermoplastic resin belonging to the family of polyester, meaning that it contains ester groups in its main polymer chain (Olam). PET is ubiquitous in our daily lives, from storage for food and beverages to raw materials to make synthetic textiles (Olam; Smelik). Conventionally, PET is produced through the polymerization between an acid, terephthalic acid (TPA), and an alcohol, ethylene glycol (EG) (Olam). The first step is known as esterification: these two components are heated together, forming oligomers and water (Olam). The temperature is kept at 220-280 degrees Celsius to remove the excess water and to dissolve TPA so it can react more readily with EG, due to its low solubility (Olam). Catalysts like amines can be added during this step to accelerate the process (Olam). Next, temperature is increased while pressure is decreased to distill the excess EG (Olam). After this, PET is formed and can be shaped through various molding methods such as injection molding (Czepiel et al.). Due to the aliphatic nature of EG and the aromatic nature of TPA, PET is a semicrystalline plastic that exhibits both crystalline and amorphous regions in its structure, and because of the rigidity of its benzene rings, the plastic has slow crystallization speed, increasing the energy (heat) and time for it to form its semicrystalline structure (Olam).

Due to the similar structural and functional characteristics of its key residues, suckerin 12 (S12) may serve as an alternative to polyethylene terephthalate (PET). One defining similarity between S12 and PET is their alternating structures. S12, as it is a block copolymer, is formed by crystalline β-sheet structures while being flanked by amorphous regions with random coils (Hershewe et al.). This results from the properties of the key residues in the protein. The GGY motifs confer flexibility, as the R group for G is simply a hydrogen atom, allowing for more freedom to move around, while A and H residues form hydrogen bonds that form β-sheets (Hershewe et al.). Correspondingly, PET, being an aliphatic-aromatic plastic, leads to a semi-crystalline structure. The integration of both structural features grants mechanical strength along with flexibility (Olam). Furthermore, PET is a thermoplastic that is able to be melted and reformed multiple times (Olam). PET also alternates between amorphous and semi-crystalline depending on the thermal environment and processing conditions. When experiencing rapid cooling, the PET chains do not have time to arrange themselves in an orderly manner (due to benzene rings), resulting in the plastic being largely amorphous (Olam). However, when there is time for slow cooling, PET has more time for its chains to form the crystalline structures (Olam). Interestingly, this property is also reflected in S12, which also exhibits thermoplastic properties, in that they can be melted and reshaped while experiencing no or minimal amount of mechanical loss (Hiew and Miserez). Research shows that when SRT is heated in water, it becomes more flexible, and after the heat is removed, the process reverses. In addition, this experiment was repeated with crushed SRT with a small amount of water, and after reaching a phase transition, the mixture was able to be processed into other shapes without seeing any significant changes in the elastic modulus of SRT (Hiew and Miserez). Moreover, S12 was shown to behave differently at varying pH levels, reflecting its pH dependence (Sun and Ding) . Due to its short sequence and the ability to display key features embodied by S12 (aggregation into β-structures and pH dependence), A1H1 of the M1 module is frequently used in experiments to understand the properties of S12 (Sun and Ding) . Previous studies reveal that, at low pH (<6.3), histidine (H) residues that are prone to change in pH levels become protonated (gain positive charge), and since histidine is so prevalent in the peptide, it causes electrostatic repulsion between the residues, effectively “ripping” the β-sheets apart, reducing the stability of the structure (Sun and Ding) . This destabilization process, however, can be terminated by adding salts to neutralize the solution, as demonstrated by adding chloride salts in both water and acidic environments, where S12 retained its β-sheet structure (Hershewe et al.). This phenomenon also occurred post-S12 acidification; when chloride salts were added to the acidified S12, it was able to restore destabilized β-sheet content. This was mostly due to the chloride salts neutralizing the protonated histidine residues that led to the conformational change (Hershewe et al.). This revealed the possibility of controllable self-assembly of S12; by changing the conditions of its surrounding solvent, the β-sheets could be folded or unfolded, affecting the functions of the protein.

Comparative Analysis of Beta-Sheet Crystals in MaSp1 and Crystalline Domains in Nylon 6,6

A comparison between major ampullate spidroin 1 (MaSp1) and nylon 6,6 demonstrates how repetitive amino acid sequences containing beta-sheet structures and flexible regions can compete with, and in some aspects surpass, a polyamide synthesized from petroleum derivatives. Spider silk has long been of much interest due to its versatility compared to synthetic fibers. A single spider can create up to seven types of silk, each with different purposes, such as capturing prey, egg protection, mobilization, etc (Scheller and Conrad). Of all these silks, the one that attracts the most attention is dragline/major ampullate (MA) silk, due to its high tensile strength and toughness (Lawrence et al.). This type of silk is composed of two proteins, major ampullate spidroin 1 (MaSp1) and major ampullate spidroin 2 (MaSp2), with a ratio of around 3:2, respectively (Y. Zhang et al.). MaSp1 is a large protein constructed by around 3000 to 4000 AAs (Y. Zhang et al.). It has 3 core sections in its primary sequence: a non-repetitive N-terminus domain (NR-NTD), a core extensive repetitive sequence, and a non-repetitive C-terminus domain (NR-CTD) (Y. Zhang et al.; Guessous et al.). The non-repetitive N and C-termini have approximately 100 residues that facilitate the formation of silk, and both are key in the self-assembly of MaSp1's structure; however, the repetitive section of the MaSp1 protein dominates the majority of the sequence and resulting function (Y. Zhang et al.). Unlike MaSp2, the core sequence of MaSp1 is characterized by poly-A motifs (3-8 consecutive alanine residues), GGX motifs (X=A, Q, or Y), GX motifs (X=Q, A, or R), and GSQ motifs. These motifs are organized in a pattern where a glycine-rich motif is linked by a poly-A motif, and these patterns are all usually 34-36 AAs long (Y. Zhang et al.). These AA compositions lead to a secondary structure that is rich in α-helices, β-sheets, and β-turns. Primarily, the polyA residues (polyA and GA motifs) play a crucial role in forming parallel and anti-parallel β-sheets (Y. Zhang et al.; Guessous et al.) . Research shows that glycine residues (GGA) also contribute to the formation of these structures (Y. Zhang et al.; Guessous et al.; Jenkins et al.); however, the majority of these residues form helical structures that mimic the type II 31-helical structure of poly(Gly) (Y. Zhang et al.). This helical structure is oriented along the silk fiber, allowing for mechanical strength, as any stress will dissipate along the fiber (Holland et al.). Additionally, an LGXQ motif is identified to be involved in a type-I turn structure. These polyA motifs play an important role in the tertiary structure of the protein, where the side chains (methyl group) of alanine residues contribute to stabilizing the overall structure (Holland et al.). Since MaSp1 spends the majority of its time in an aqueous environment, the hydrophobic parts of the protein begin to fold inwards, including alanine residues. This guides the formation of the β-sheets, which are at the core and stacked closely together (Holland et al.). Upon this stacking, the methyl groups of alanine residues protrude into the empty spaces created by the backbone of opposing sheets. This protrusion results in extremely tight packing of the protein, resulting in a rigid and strong structure. This protrusion may not be possible if another hydrophobic AA replaced the alanine residue due to the simplicity of alanine's side chain, which prevents steric clashes with any larger side chain. This efficient packing at the molecular level forms the foundation for the larger quaternary assemblies, where multiple β-sheet-rich domains organize into robust supramolecular structures. Additionally, NR-NTD, which is pH-sensitive, plays a crucial role in

establishing initial contact with other protein units to form the quaternary structure (Guessous et al.). In the spider's major ampullate silk gland, the pH is relatively acidic, and the protein is highly concentrated in a soluble state (Landreh et al.). When spinning occurs, the protein travels from the gland to the duct, where the pH of the environment changes from 6.9 to 6.3, and drives the dimerization of two NTDs that connect two MaSp1 proteins. This causes the transition from random coils to the formation of β-sheets rich fibers. This allows for further stabilization (protonates acidic AAs) and interactions between the secondary structures of the two proteins, assembling the strong fiber (Landreh et al.).

After exploring the structural and mechanical features of MaSp1, it is interesting to draw comparisons to synthetic polymers like nylon 6,6. Polyamide polymers, otherwise known as nylon, are a group of thermoplastics that are linear polymers linked by amide groups (Grishanov). Nylon has wide applications in textiles, electronics, and materials for automobile parts (McNeeley and Liu). The most common types of nylon are nylon 6 (6 represents the number of carbon in the monomer) or nylon 6,6 (6 carbon atoms in each of the two monomers) (Grishanov). The two monomers that compose nylon 6,6 are adipic acid and hexamethylene diamine (HMDA) (McNeeley and Liu). Similar to many other synthetic polymers, nylon 6,6 has a long polymer chain, contains both crystalline and amorphous regions, and can interact with neighboring chains through van der Waals forces and hydrogen bonds (Al-Shawabkeh). These fibers are created through a process called melt-spinning, a common process in producing thermoplastics (Grishanov). The polymers are first sent into a spinneret where they are melted and pressurized (Naeimirad et al.). Next, the polymers go through a spin pack, which filters the polymer and forms filaments of the desired amount and cross-section. The processed polymers then leave the machine and enter water to cool down and solidify (Naeimirad et al.). The solidified fibers are then subjected to solid-state drawing, which involves stretching the extruded polymers at a draw ratio to align the molecules, achieving a higher crystalline state (Mai et al.). This process can significantly increase the elongation and toughness compared to the alternative polymer.

The shared structural and functional features of MaSp1 suggest its potential as an alternative to nylon 6,6. The reason that MaSp1 has been at the center of scientific research is due to its extraordinary mechanical properties that match or even surpass high-tech synthetic fibers and its biodegradable nature. The key to the spidroin's high mechanical properties is a result of its repetitive core of the protein, which constitutes around 90% of the structure (Guessous et al.). These repetitive cores form the basis of β-sheets, which are formed by hydrogen bonds and reinforced by hydrophobic interactions and van der Waals forces (Cohen and Eisenbach). Any stress put on this structure will need to overcome the combination of all these forces to successfully "break" the structure, resulting in high mechanical strength. Furthermore, the β-sheet domains act as crosslinkers or "fillers" for the amorphous regions, which allows for an efficient way to transfer load throughout the protein (Cohen and Eisenbach). With the high content of these proteins in major ampullate (MA) silk, it can achieve a tensile strength of 1 or more GPa, which rivals many synthetic polymers (Lawrence et al.). In

comparison, nylon 6,6 is also extremely strong among synthetic polymers, due to its drawing modification after melt-spinning and the crosslinkages between nylon polymers via hydrogen bonding (Al-Shawabkeh). As a result, nylon 6,6 composites can have a tensile strength of up to 79.5 MPa (Bhavya and Savitha). For both polymers, chain length is a prominent factor that may directly dictate the final mechanical strength. In particular, research shows that in recombinant MaSp1 (re-MaSp1) proteins, there is a linear relationship between amino acid length and mechanical properties (You et al.). For example, there is a positive correlation between the maximum stress (MPa) and the number of AAs in re-MaSp1 proteins.

Discussion

This review aims to clarify how specific amino acid sequences in structural proteins give rise to plastic-like properties and how these properties allow these proteins to be viewed as future alternatives for plastics; we compared glycinin 11s to polyethylene (PE), suckerin-12 to polyethylene terephthalate (PET), and major ampullate spidroin 1 to nylon 6,6. The analysis revealed a shared molecular design between the petroleum-derived polymers and their biodegradable counterparts: the presence of semi-crystalline structures. Specifically, the β-sheets that form the crystalline sections in proteins, which supply them with mechanical and tensile strength, and the amorphous regions created by flexible amino acids like glycine, which offer flexibility, mirror the semi-crystalline conformation of synthetic plastics. These repetitive structures suggest that the arrangement and pattern of these alternating structures are key to which biological and petroleum-based polymers achieve their functions. However, having this common molecular structure established, it is important to consider the advantages and limitations that protein-based alternatives have compared to conventional plastics.

In general, proteins, due to their origin as naturally occurring polymers, offer several inherent advantages that allow them to compete with conventional plastics: biodegradability and tunability.

Since the rise of plastics in the 19th century, they have become an increasing threat to the environment. More recently, microscopic pieces of plastics known as microplastics have been discovered in nearly every part of the biosphere. These microplastics are not only small (< 5 mm), but also non-degradable, able to last in our environment for a long time. The accumulation of these plastics stems from two main sources: Primary and Secondary microplastics (Ziani et al.). Primary microplastics are manufactured intentionally for specific uses, such as microbeads used in the cosmetics industry. Secondary microplastics, on the other hand, result from the fragmentation of larger pieces of plastic and are far more prevalent than primary microplastics. Together, these plastics have detrimental effects on our ecosystems, acting as physical hazards that can lead to ingestion and entanglement in marine organisms, and chemical vectors for toxic additives that can lead to inflammation and disruptions in metabolism (Ziani et al.). These pervasive environmental and physiological risks of plastics highlight the urgent need for biodegradable protein alternatives. In proteins, the amino acids that compose them are connected via peptide bonds. These bonds are susceptible to enzymatic hydrolysis by proteases produced by

the microorganisms in natural environments. These proteases depolymerize protein-based materials as the polypeptide is cleaved into smaller peptides or amino acids. These fragments can be integrated into the environment via nutrient cycles or be reused to create new proteins. For example, in soy protein isolate films, which glycinin 11s constitute a significant structural component of, demonstrate remarkable biodegradability. When films are buried in the soil, the hydrophilic amino acids begin to facilitate water absorption, leading to swelling. As this happens, the peptide linkages between amino acids are exposed to proteolytic enzymes secreted by microorganisms in the soil, allowing the proteins to be hydrolyzed into smaller fragments (González et al.). Furthermore, the biodegradation rate is tunable by increasing or decreasing cross-linkages in the structure: increased cross-linkages can reduce swelling and enzyme diffusion, slowing down the degradation rate (González et al.). Therefore, protein-based materials are able to alleviate environmental concerns posed by petroleum-based plastics.

Other than biodegradability, protein-based materials offer more precise tunability over structure-to-function relationships than that of plastics. In conventional plastics, the properties of the material, such as mechanical strength, stiffness, and thermal behavior, arise from molecular weight, crystallinity, copolymer composition, and additive incorporation (Sperling; Odian). To alter or refine these properties, bulk chemical processes are used. For example, by restricting chain mobility through methods like controlled cooling or drawing, it can increase the crystallinity of the material, thus enhancing its overall tensile strength (Sperling; Odian). In contrast, adding plasticizers or fillers can increase flexibility and impact resistance (Sperling). Additionally, copolymerization allows for further modification by incorporating two or more different monomers in a single polymer chain. The extent of the modification depends on the reaction conditions and monomer ratios; for example, by adding a monomer that may disrupt packing, its presence can lower crystallinity levels, producing more elastic and flexible behavior (Odian). Still, these modifications are done on a macroscopic level rather than precisely controlling the molecular polymer structure. The tunability of proteins, however, is more flexible and can be done on a fundamental level, enabling complete control over their molecular structures. Returning to Anfinsen's foundational principles of biochemistry, where the sequence of amino acids determines structure and thus function (Kresge et al.), this idea is fundamental to the tunability of proteins. Since amino acids constitute a protein, by changing these units, a protein's structure can be altered as well. Spider silk proteins, for example, specifically major ampullate spidroins (MAS), demonstrate high tensile strength while also demonstrating impressive flexibility (Heim et al.). The high tensile strength emerges from the amount and tightness of packing of the β-sheets (remaining as polyproline-II helix-like structures depending on the environment) that are mainly clustered in the repetitive domain of the protein, and the glycine-rich regions lead to amorphous domains for elasticity (Heim et al.). These sequences of crystalline and amorphous sections determine molecular orientation and structure, allowing tuning of the sequence to achieve different macroscopic effects. Research has discovered that the number of repeats, the modular motif arrangement, and codon usage for recombinant expression of spider-silk genes can be tuned through cloning systems that assemble oligonucleotides into

synthetic spider-silk genes (Heim et al.). Through these systems, scientists were able to mimic the silk properties of many spiders, including the dragline silk of Araneus diadematus and the flagelliform silk proteins of Nephila clavipes (Heim et al.).

However, current protein production is still underdeveloped in many aspects and fails to create high-quality protein-based materials efficiently. Most prominently, finding a reliable protein source may prove to be difficult for many materials. These include the specialized structural proteins found in spidroins and squid teeth. Harvesting them directly from their hosts is not practical, as these proteins only occur in small amounts in their natural hosts; thus, obtaining these proteins naturally yields amounts that are far from enough to be utilized on an industrial scale. To increase the production efficiency, scientists have turned their attention to recombinant expression systems, where the genes of these proteins are expressed in non-native host organisms to produce the corresponding proteins (Lazaris et al.). For example, fragments of MaSpI and MaSpII have been recombinantly expressed in cultured mammalian cells (Lazaris et al.); similarly, SRT proteins are expressed in *Escherichia Coli* (*E. Coli*), as they are readily available and it has a well-understood genetic system (Pena-Francesch and Demirel). However, once scaled up to an industrial level, this method of production becomes costly and time-consuming compared to plastic production. For SRT proteins, the yield for 80L of fermenters was only 1g/L (one gram produced per liter of culture); accounting for purification, the estimated cost was around $100 per kilogram (Pena-Francesch and Demirel). For comparison, LDPE's reference value is around $1 per kilogram (Hernández et al.).

Furthermore, another major limitation for protein-based alternatives is their sensitivity to different environmental conditions. Although being sensitive to the environment may be beneficial at times, as seen in the case of biodegradation, it can also weaken the material during usage. For example, soy protein isolates, which contain many hydrophilic amino acids, are extremely sensitive to polar environments; thus, films constructed by these proteins are vulnerable to water and moisture, as water molecules can interfere with the film structure and cause swelling (González et al.). Likewise, suckerin-12's sensitivity to pH levels also shows protein-based materials' instability upon varying environmental conditions (Sun and Ding; Hershewe et al.).

To overcome these limitations, one future direction to focus on is improving production efficiency. To produce protein-based materials, recombinant expression systems must operate to produce larger quantities of structural proteins and with reduced inputs. This requires the non-native host organism to be compatible with the demands of the recombinant protein. If the host cannot supply enough amino acids for production, it can result in a low yield (Scheller and Conrad). Moreover, this system needs to maintain consistent accuracy and quality with its products. This is important as structural proteins' properties rely heavily on amino acid sequences and the amount of repeated motifs. Any inconsistencies in the production may directly impact the material's final performance. Another aspect to focus on is the balance between durability and biodegradability. It was mentioned before that an advantage of protein-based materials is their ability to self-depolymerize under certain environmental conditions; however,

this sensitivity can influence the stability of the material. Thus, it is key to achieve a balance between the amount of cross-linkages, β-sheet content, hydrophobicity, and other processing conditions that allow the material to remain stable during use, but still biodegradable after disposal. The future of protein-based materials requires not only the identification of useful protein properties, but also the production and processing methods that can replicate those properties on an industrial scale.

Conclusion

In conclusion, the similarities regarding molecular structure and function between structural proteins and conventional plastics suggest proteins' potential as a possible alternative to petroleum-based products. The three comparisons in this paper show that the potential comes from proteins' amino acid sequence, which determines their structure and, in turn, their function. Therefore, this connection between sequence and function of proteins and plastics shows that creating sustainable alternatives is much more complex than simply finding a natural material to use as an alternative; instead, future sustainable alternatives must be designed at the molecular level, with specific sequences altered to fit a purpose.

Works Cited

Adachi, Motoyasu, et al. “Crystal Structure of Soybean 11S Globulin: Glycinin A3B4 Homohexamer.” Proceedings of the National Academy of Sciences, vol. 100, no. 12, June 2003, pp. 7395–400, https://doi.org/10.1073/pnas.0832158100.

Al-Shawabkeh, Ali F. “Optoelectronic Investigation and Spectroscopic Characteristics of Polyamide-66 Polymer.” E-Polymers, vol. 22, no. 1, 2022, pp. 858–69, https://doi.org/10.1515/epoly-2022-0078.

Amjadi, Mohammad, and Ali Fatemi. “Tensile Behavior of High-Density Polyethylene Including the Effects of Processing Technique, Thickness, Temperature, and Strain Rate.” Polymers, vol. 12, no. 9, Sept. 2020, p. 1857, https://doi.org/10.3390/polym12091857.

Berg, Paul. “Fred Sanger: A Memorial Tribute.” Proceedings of the National Academy of Sciences of the United States of America, vol. 111, no. 3, Jan. 2014, pp. 883–84, https://doi.org/10.1073/pnas.1323264111.

Bhavya, V., and M. Savitha. “Mechanical and Thermal Properties of Nylon 66/Flyash Composites: Effect of Flyash.” International Journal of Scientific Research in Engineering and Management, May 2023, https://ijsrem.com/download/mechanical-and-thermal-properties-of-nylon-66-flyash-composites-effect-of-flyash/.

Calinsky, Rivka, and Yaakov Levy. “Histidine in Proteins: pH-Dependent Interplay between π–π, Cation–π, and CH–π Interactions.” Journal of Chemical Theory and Computation, vol. 20, no. 15, July 2024, pp. 6930–45, https://doi.org/10.1021/acs.jctc.4c00606.

Chen, Junfeng, et al. “Intramolecularly Cross-Linked Polymers: From Structure to Function with Applications as Artificial Antibodies and Artificial Enzymes.” Accounts of Chemical Research, vol. 53, no. 7, 2020, pp. 1401–10, https://doi.org/10.1021/acs.accounts.0c00178.

Cho, Seung Yong, and Chul Rhee. “Mechanical Properties and Water Vapor Permeability of Edible Films Made from Fractionated Soy Proteins with Ultrafiltration.” LWT - Food Science and Technology, vol. 37, no. 8, Dec. 2004, pp. 833–39, https://doi.org/10.1016/j.lwt.2004.03.009.

Chremos, Alexandros, and Jack F. Douglas. “Influence of Branching on the Configurational and Dynamical Properties of Entangled Polymer Melts.” Polymers, vol. 11, no. 6, June 2019, p. 1045, https://doi.org/10.3390/polym11061045.

Cohen, Noy, and Claus D. Eisenbach. “Molecular Mechanics of Beta-Sheets.” ACS Biomaterials Science & Engineering, vol. 6, no. 4, Apr. 2020, pp. 1940–49, https://doi.org/10.1021/acsbiomaterials.9b01983.

Collins, Francis S., and Leslie Fink. “The Human Genome Project.” Alcohol Health and Research World, vol. 19, no. 3, 1995, pp. 190–95.

Czepiel, Mateusz, et al. “Advanced Injection Molding Methods: Review.” Materials, vol. 16, no. 17, 2023, p. 5802, https://doi.org/10.3390/ma16175802.

dos Santos, Everton Ranny Ferreira. Nanostructure Materials as Catalysts for the Degradation of Polyolefins. 2018, Universidade de Lisboa, Instituto Superior Técnico, PhD thesis, https://scholar.tecnico.ulisboa.pt/api/records/fZSEu8nulGKwRmk7oxubk8sxt5rcK848ownH/file/157a1cb14030372692cce175d53579ba48ec82614156ab95a99699452ac35441.pdf.

Emblem, A. "Plastics Properties for Packaging Materials." Packaging Technology, edited by Anne Emblem and Henry Emblem, Woodhead Publishing, 2012, pp. 287–309, https://doi.org/10.1533/9780857095701.2.287.

González, Agustín, et al. "Cross-Linked Soy Protein as Material for Biodegradable Films: Synthesis, Characterization and Biodegradation." Journal of Food Engineering, vol. 106, no. 4, Oct. 2011, pp. 331–38, https://doi.org/10.1016/j.jfoodeng.2011.05.030.

Grishanov, S. "Structure and Properties of Textile Materials." Handbook of Textile and Industrial Dyeing, edited by M. Clark, vol. 1, Woodhead Publishing, 2011, pp. 28–63, https://doi.org/10.1533/9780857093974.1.28.

Guessous, Ghita, et al. "Disentangling the Web: An Interdisciplinary Review on the Potential and Feasibility of Spider Silk Bioproduction." ACS Biomaterials Science & Engineering, vol. 10, no. 9, Aug. 2024, pp. 5412–38, https://doi.org/10.1021/acsbiomaterials.4c00145.

Heim, Markus, et al. "Spider Silk: From Soluble Protein to Extraordinary Fiber." Angewandte Chemie International Edition, vol. 48, no. 20, May 2009, pp. 3584–96, https://doi.org/10.1002/anie.200803341.

Hernández, Borja, et al. "Techno-Economic and Life Cycle Analyses of Thermochemical Upcycling Technologies of Low-Density Polyethylene Waste." ACS Sustainable Chemistry & Engineering, vol. 11, no. 18, 2023, pp. 7170–81, https://doi.org/10.1021/acssuschemeng.3c00636.

Hershewe, Jasmine M., et al. "Characterizing and Controlling Nanoscale Self-Assembly of Suckerin-12." ACS Synthetic Biology, vol. 9, no. 12, Dec. 2020, pp. 3388–99, https://doi.org/10.1021/acssynbio.0c00442.

Hiew, Shu Hui, and Ali Miserez. "Squid Sucker Ring Teeth: Multiscale Structure–Property Relationships, Sequencing, and Protein Engineering of a Thermoplastic Biopolymer." ACS Biomaterials Science & Engineering, vol. 3, no. 5, May 2017, pp. 680–93, https://doi.org/10.1021/acsbiomaterials.6b00284.

Holland, Gregory P., et al. "Determining Secondary Structure in Spider Dragline Silk by Carbon−Carbon Correlation Solid-State NMR Spectroscopy." Journal of the American Chemical Society, vol. 130, no. 30, July 2008, pp. 9871–77, https://doi.org/10.1021/ja8021208.

Hu, Zhaodong, et al. "Impacts of Industrial Modification on the Structure and Gel Features of Soy Protein Isolate and Its Composite Gel with Myofibrillar Protein." Foods, vol. 12, no. 10, May 2023, p. 1982, https://doi.org/10.3390/foods12101982.

Jenkins, Janelle E., et al. “Characterizing the Secondary Protein Structure of Black Widow Dragline Silk Using Solid-State NMR & X-Ray Diffraction.” Biomacromolecules, vol. 14, no. 10, Oct. 2013, pp. 3472–83, https://doi.org/10.1021/bm400791u.

Kehinde, A. J., et al. “Emerging Issues in the Mechanisms of High Pressure Free Radical Ethylene Polymerization: A Review.” American Journal of Polymer Science, vol. 2, no. 5, 2012, https://doi.org/10.5923/j.ajps.20120205.03.

Klute, C. H., and P. J. Franklin. “The Permeation of Water Vapor through Polyethylene.” Journal of Polymer Science, vol. 32, no. 124, Oct. 1958, pp. 161–76, https://doi.org/10.1002/pol.1958.1203212413.

Kresge, Nicole, et al. “The Thermodynamic Hypothesis of Protein Folding: The Work of Christian Anfinsen.” Journal of Biological Chemistry, vol. 281, no. 14, Apr. 2006, pp. e11–13, https://doi.org/10.1016/S0021-9258(19)56522-X.

Landreh, Michael, et al. “A pH-Dependent Dimer Lock in Spider Silk Protein.” Journal of Molecular Biology, vol. 404, no. 2, Nov. 2010, pp. 328–36, https://doi.org/10.1016/j.jmb.2010.09.054.

Lawrence, Barbara A., et al. “Molecular and Mechanical Properties of Major Ampullate Silk of the Black Widow Spider, Latrodectus h Esperus.” Biomacromolecules, vol. 5, no. 3, May 2004, pp. 689–95, https://doi.org/10.1021/bm0342640.

Lazaris, Anthoula, et al. “Spider Silk Fibers Spun from Soluble Recombinant Silk Produced in Mammalian Cells.” Science, vol. 295, no. 5554, 2002, pp. 472–76, https://doi.org/10.1126/science.1065780.

Lin, Song, et al. “The Effect of Waste Plastic on Environmental Degradation: A Corporate Perspective.” Polish Journal of Environmental Studies, vol. 34, no. 1, Jan. 2025, pp. 203–11, https://doi.org/10.15244/pjoes/185999.

Liu, Yuying, et al. “Molecular Insights into the Self-Assembly of Block Copolymer Suckering Polypeptides into Nanoconfined β-Sheets.” Small, vol. 18, no. 34, Aug. 2022, p. e2202642, https://doi.org/10.1002/smll.202202642.

Mai, Fang, et al. “The Influence of Solid-State Drawing on Mechanical Properties and Hydrolytic Degradation of Melt-Spun Poly(Lactic Acid) (PLA) Tapes.” Fibers, vol. 3, no. 4, 2015, pp. 523–38, https://doi.org/10.3390/fib3040523.

Maruyama, Nobuyuki, et al. “Structure–Physicochemical Function Relationships of Soybean Glycinin at Subunit Levels Assessed by Using Mutant Lines.” Journal of Agricultural and Food Chemistry, vol. 52, no. 26, 2004, pp. 8197–8201, https://doi.org/10.1021/jf048786y.

McNeeley, Adam, and Y. A. Liu. “Assessment of Nylon-66 Depolymerization for Circular Economy: Kinetic Modeling, Purification, and Sustainable Design.” Industrial & Engineering Chemistry Research, vol. 64, no. 8, Feb. 2025, pp. 4179–203, https://doi.org/10.1021/acs.iecr.4c04411.

Naeimirad, Mohammadreza, et al. “A Review on Melt-Spun Biodegradable Fibers.” Sustainability, vol. 15, no. 19, 2023, p. 14474, https://doi.org/10.3390/su151914474.

Odian, George. Principles of Polymerization. 4th ed., Wiley, 2004.

Olam, Mikail. “PET: Production, Properties and Applications.” Advances in Materials Science Research, edited by Maryann C. Wythers, vol. 48, Nova Science Publishers, 2021, pp. 131–62.

Pena-Francesch, Abdon, and Melik C. Demirel. “Squid-Inspired Tandem Repeat Proteins: Functional Fibers and Films.” Frontiers in Chemistry, vol. 7, 2019, article 69, https://doi.org/10.3389/fchem.2019.00069.

Perutz, M. F., et al. “Aggregation of Proteins with Expanded Glutamine and Alanine Repeats of the Glutamine-Rich and Asparagine-Rich Domains of Sup35 and of the Amyloid Beta-Peptide of Amyloid Plaques.” Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 8, Apr. 2002, pp. 5596–600, https://doi.org/10.1073/pnas.042681599.

Pey, Angel L., et al. “Engineering Proteins with Tunable Thermodynamic and Kinetic Stabilities.” Proteins, vol. 71, no. 1, Apr. 2008, pp. 165–74, https://doi.org/10.1002/prot.21670.

Peng, I. C., et al. “The Physicochemical and Functional Properties of Soybean 11S Globulin—A Review.” Cereal Chemistry, vol. 61, no. 6, 1984, pp. 480–90.

Pham, Nga Thi-Hong. “Characterization of Low-Density Polyethylene and LDPE-Based/Ethylene-Vinyl Acetate with Medium Content of Vinyl Acetate.” Polymers, vol. 13, no. 14, July 2021, p. 2352, https://doi.org/10.3390/polym13142352.

Scheller, Jürgen, and Udo Conrad. “Plant-Based Material, Protein and Biodegradable Plastic.” Current Opinion in Plant Biology, vol. 8, no. 2, 2005, pp. 188–96, https://doi.org/10.1016/j.pbi.2005.01.010.

Shamsir, Mohd S., and Andrew R. Dalby. “β-Sheet Containment by Flanking Prolines: Molecular Dynamic Simulations of the Inhibition of β-Sheet Elongation by Proline Residues in Human Prion Protein.” Biophysical Journal, vol. 92, no. 6, Mar. 2007, pp. 2080–89, https://doi.org/10.1529/biophysj.106.092320.

Sharma, A., and S. Sharma. “Effect of Nitrogen and Sulphur Nutrition on Storage Protein Quality in Soybean [Glycine Max (L.) Merrill].” Journal of Applied and Natural Science, vol. 10, no. 1, Mar. 2018, pp. 296–300, https://doi.org/10.31018/jans.v10i1.1620.

Shi, Baoshan, et al. “Measurement and Analysis on Tensile Strength and Crystallinity of Vibrating Extruded High-Density Polyethylene (HDPE) Sheet.” Advanced Materials Research, vols. 291–294, July 2011, pp. 561–64, https://doi.org/10.4028/www.scientific.net/AMR.291-294.561.

Singh, Amandeep, et al. “Structural and Functional Analysis of Various Globulin Proteins from Soy Seed.” Critical Reviews in Food Science and Nutrition, vol. 55, no. 11, Sept. 2015, pp. 1491–502, https://doi.org/10.1080/10408398.2012.700340.

Smelik, Anneke. “Polyester: A Cultural History.” Fashion Practice, vol. 15, no. 2, May 2023, pp. 279–99, https://doi.org/10.1080/17569370.2023.2196158.

Sperling, L. H. Introduction to Physical Polymer Science. 4th ed., Wiley, 2006.

Stollar, E. James, and David P. Smith. “Uncovering Protein Structure.” Essays in Biochemistry, vol. 64, no. 4, Oct. 2020, pp. 649–80, https://doi.org/10.1042/EBC20190042.

Sun, Yunxiang, and Feng Ding. “Thermo- and pH-Responsive Fibrillization of Squid Suckerin A1H1 Peptide.” Nanoscale, vol. 12, no. 11, Mar. 2020, pp. 6307–17, https://doi.org/10.1039/c9nr09271d.

Wang, Yufei, and Haifeng Qian. “Phthalates and Their Impacts on Human Health.” Healthcare, vol. 9, no. 5, May 2021, p. 603, https://doi.org/10.3390/healthcare9050603.

Yang, Yuhui, et al. “The Degree of Crystallinity Exhibiting a Spatial Distribution in Polymer Films.” European Polymer Journal, vol. 107, Oct. 2018, pp. 303–07, https://doi.org/10.1016/j.eurpolymj.2018.08.041.

You, Zhengying, et al. “Extraordinary Mechanical Properties of Composite Silk Through Hereditable Transgenic Silkworm Expressing Recombinant Major Ampullate Spidroin.” Scientific Reports, vol. 8, no. 1, Oct. 2018, p. 15956, https://doi.org/10.1038/s41598-018-34150-y.

Zhang, Haimei, editor. “Building Plastic.” Building Materials in Civil Engineering, Woodhead Publishing, 2011, pp. 289–423, https://doi.org/10.1533/9781845699567.289.

Zhang, Yang, et al. “The Molecular Structures of Major Ampullate Silk Proteins of the Wasp Spider, Argiope Bruennichi: A Second Blueprint for Synthesizing de Novo Silk.” Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, vol. 164, no. 3, Mar. 2013, pp. 151–58, https://doi.org/10.1016/j.cbpb.2012.12.002.

Ziani, Khaled, et al. “Microplastics: A Real Global Threat for Environment and Food Safety: A State of the Art Review.” Nutrients, vol. 15, no. 3, Jan. 2023, p. 617, https://doi.org/10.3390/nu15030617.

Evaluating Hand-Tracking Failures in Sign-Language Videos: The Roles of Skin Tone and Video Conditions By Kshiteej Herode

Abstract

Sign-language recognition systems begin with a hand-tracking step that locates the signer's hands and marks the positions of their joints. Every later step of the system depends on this initial detection. Whereas existing work has focused on the accuracy of the final recognition output, this study analyzed the hand tracking directly. Specifically, Google's MediaPipe HandLandmarker was tested to see whether it fails more often for hands with a darker estimated skin tone once video conditions were taken into account. Fifty publicly available sign-language clips from the WLASL-100 subset were processed, from which every third frame was kept. Every signer's skin tone was estimated, and based on the estimated skin tone of the signer, videos were placed in the lighter or darker group. Each kept frame was scored for hand detection, brightness, and motion; resolution was scored once per video. The data showed that the darker group had a higher detection rate, with hands detected in an average of 71.9% of frames compared to 63.1% in the lighter group. Alongside this, brightness and resolution also had a positive relationship with detection. However, there was a cluster of 11 unusual videos that had high brightness and almost perfect detection, but the lowest resolution in the sample. When those videos were excluded, the skin-tone effect disappeared, but brightness and resolution were still important factors that affected the detection rate. In the end, this study did not find evidence that skin tone affects hand detection, but it did find evidence that brightness and resolution were consistently linked to improved detection rates.

1. Introduction

Sign languages are complete forms of communication that have their own grammar and vocabulary. They are "spoken" through the shape of the hand, the positions of the fingers, where the hand is placed, how the hand moves, and the expression on the signer's face. For millions of people, specifically those who are deaf or hard of hearing (DHH), sign languages are the primary means of communication. In fact, the World Federation of the Deaf estimates that more than 70 million deaf people use sign languages (Bragg et al. 16), demonstrating their critical role in communication. The importance of sign languages has led to the development of advanced systems for recognizing and interpreting them. Systems that can read signed video, such as ones that can translate between signs and spoken languages or ones that can caption signs, help improve communication tools for DHH communities (Bragg et al.). Most modern sign-recognition systems are built as pipelines, meaning the output of one stage is a part of the input to the next. The first stage is almost always the same: looking at each frame of a video and marking the locations of joints, such as wrists, knuckles, and fingertips, as a set of points called keypoints. However, if this step fails, everything downstream fails with it. If a signer's hands are detected poorly, every tool built on top receives inaccurate data, meaning poorer captions and translation. That failure happens often; MediaPipe, a popular tool for hand tracking, misses over

50% of the hands present in WLASL (Roh et al. 323), the largest public dataset of word-level American Sign Language clips (Li et al.). The cascading impact of these failures is huge when it is considered how often they happen.

Existing research on sign-language systems has focused on the final output, measuring whether the recognized word is correct. While this work is important to improve the accuracy of sign-language systems, it does not answer whether the issue is the initial hand recognition, the deciphering algorithm, or any one of the many steps in the pipeline. This study begins at the start of that pipeline, asking whether hand detection is the issue, and more specifically, whether hand detection is even across signers of different skin tones. The hypothesis was that MediaPipe would detect hands less often for darker-skinned signers. To test this, the study was designed to separate skin tone from recording conditions. Fifty publicly available clips from WLASL-100 were processed through Google's MediaPipe HandLandmarker. Every third frame was processed and scored for whether a hand was detected, brightness, and motion. Each video's resolution was also recorded. Once skin tone was estimated, the 50 videos were split into two groups: lighter and darker skin tone. Then, a logistic regression was used to check whether the skin-tone group still mattered after brightness, resolution, and motion were accounted for, along with a subsequent check to see whether the result held up after setting aside a small cluster of unusual videos.

2. Related Work

Work on sign language technology has mainly focused on the accuracy of the recognition phase, which is one of the later stages. Most recently, Atwell et al. found that recognition accuracy was higher for signers with lighter skin tones (5). In their study, they flagged lighting and video quality as confounding factors they were unable to fully separate from skin tone, which is the separation that this study aims to do. Because their analysis measured only the final output, it cannot say which stage in the pipeline causes the gap between the accuracy of lighter and darker skin tones. Next, working to improve sign-language recognition, Roh et al. found that MediaPipe fails to detect more than half of the hands present in WLASL. This results in a large share of data being lost, preventing the later stages from working well. The large number of failures, over 50%, warrants an investigation for any patterns, which is why this study checks whether or not the tracker fails evenly across signers of different skin tones. Furthermore, concerns about skin-tone bias are not limited to hand detection. Wilson et al. tested whether object-detection software used in self-driving detects pedestrians equally across skin tones. They found that darker-skinned pedestrians were detected less reliably. Even after accounting for other factors, such as lighting at night and scenes where the view of pedestrians was blocked, their finding still held up. Their study is a concrete example of demographic bias in detection systems. Parte et al. analyzed a face-tracking tool to determine if it functioned worse for particular demographic groups. Initially, their results did show demographic bias, but once they accounted for other factors, like head pose and resolution, gender and race effects went away while a bias

against older individuals remained. This study applies the same logic to sign-language video by checking if skin-tone affects hand detection after accounting for video conditions.

3. Methods

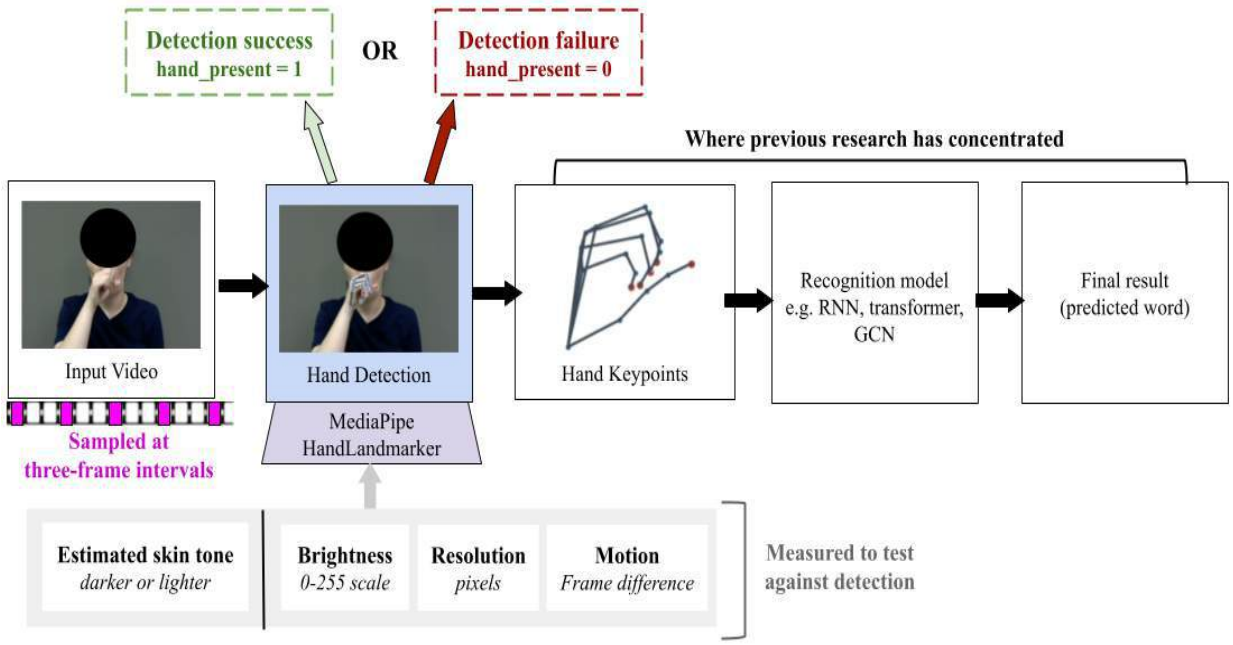


Figure 1. Hand detection in a sign-language recognition pipeline: input video is sampled at three-frame intervals, and each sampled frame passed to MediaPipe's HandLandmarker, returning either hand keypoints/landmarks (hand_present = 1) or failing to detect a hand (hand_present = 0). Brightness, motion, resolution, and estimated skin tone were measurements tested against detection

All videos were sourced from WLASL (Word-Level American Sign Language), a publicly available dataset of short clips, each depicting a single signer producing one ASL sign (Li et al.). The full detection pipeline after obtaining the sign-language clips from WLASL is shown in Figure 1. WLASL was chosen for four reasons. Firstly, it is fully public and accessible for free, allowing this analysis to be inexpensive, entirely reproducible, and require no participants. Secondly, WLASL is one of the most widely used datasets in sign-language recognition research, allowing for direct comparison of results with prior work. Furthermore, a 2024 study found that MediaPipe fails to detect over 50% of the hands in WLASL (Roh et al. 323), ensuring WLASL has an abundance of detection failures to analyze, which is what this study requires. A dataset where detection was nearly always successful would offer no variation to study; failures could not be attributed to skin tone, lighting, or any other possible factor since there would be no failures. Lastly, WLASL clips vary in brightness and resolution. This was the most important reason because it allows the separation of skin tone from recording conditions. Checking the effect of skin tone requires comparing videos with darker and lighter signers

recorded under similar conditions. On the other hand, isolating and checking the effect of conditions means comparing bright and dim clips of similar skin tone, and also comparing high-res and low-res clips of similar skin tone. The logistic regression performs exactly those comparisons.

Videos were accessed using Kaggle's command-line tool from a publicly hosted copy of the dataset in which clips were already extracted, so no scraping from the source was required. Specifically, this study drew from WLASL-100, a reduced subset containing 100 sign classes. Each sign class is a single vocabulary word, such as "hello" or "eat." Each class includes clips of many different signers producing that same sign; therefore, WLASL-100 covers 100 signs but contains roughly 2,000 clips in total. This study attempted to access the first 80 clips listed in the WLASL-100 index, of which 30 were deprecated and no longer accessible. That left 50 clips available for use. Because the clips were taken in index order and not randomly sampled, the results cannot be generalized to WLASL as a whole. This is discussed further in the limitations section.

Detecting Hands with MediaPipe

Hand detection was performed using Google's MediaPipe, a free, popular, and open-source library for detecting body keypoints in images. Specifically, its HandLandmarker component is used to obtain the coordinates of knuckles, fingertips, and other hand landmarks. MediaPipe's older, unified interface for detecting hands, face, and body pose was not used because of deprecation. For each processed frame, a binary variable, hand_present, was used to record a success (hand_present = 1), indicating that at least one hand was captured, or a failure (hand_present = 0). A failure is the outcome analyzed in this study, as it is the instance in which the tracker returned no hand. Furthermore, not every frame was processed; instead, every third frame was used because frames so close together are nearly identical. If every frame were to be used, more computing power would have been exhausted in exchange for meager additional information. In the end, 1,149 processed frames were used across the 50 videos after keeping every third frame. Detection rate was simply the proportion of processed frames scored as a success (hand_present = 1) over the total frames in the video, followed by conversion to a percentage. Each skin-tone group's detection rate was determined by averaging the individual detection rates of each video in the group.

Three measurements were made to record the video conditions. After conversion to grayscale, frame brightness was determined by the mean pixel value on a 0 (pure black) to 255 (pure white) scale. Next, motion was computed as the mean absolute pixel difference between the current frame and the previous retained frame. Since the first retained frame of each video had no previous frame, its motion value was blank and dropped from the regression accordingly. The removal of the first frame from each video left 1,099 frames that went into the logistic regression. Resolution was recorded per video instead of per frame because it stays constant throughout.

Estimating Skin Tone

Skin tone is a uniform variable across all frames, so it was estimated once per video. Because only using a single frame could provide a weak estimate, five frames from equal intervals were sampled for skin-tone estimation. Each frame was processed with an open-source skin-tone classifier, with color mode enforced on every call. As Atwell et al. found when running the same classifier, the tool tends to detect frames as black and white if color mode is not specified, which produces unusable values. The skin-tone classifier returns its output in two forms: the precise RGB values of the detected color or a categorical label based on the nearest entry on a 36-tone palette. While the skin tones were grouped into two groups later, regardless, the exact RGB values were chosen to maximize accuracy so that the determination of labels wouldn't rely on the preset ones in the tool. The tool obtains the RGB value by recording the dominant color from each of the five sampled frames. The dominant color is three numbers from 0 to 255: red, green, and blue. To make the ranking of colors from lighter to darker simpler, these three numbers were combined into a single value using the Rec. 601 luma formula: 0.299R + 0.587G + 0.114B. Each color's number is multiplied by its weight, and then all three are added. This is done for each of the five frames. Then, the five values, one from each frame, were averaged to produce a final skin-tone value per video. After all frames of all videos were processed, a partition was made at 137.4, the median value, to separate the videos into two groups of 25, lighter and darker.

Separating Skin Tone From Video Conditions

A popular statistical model for predicting a binary outcome from multiple variables simultaneously, or a logistic regression, was the central analysis of this study. The dependent variable was whether a hand was detected in a given frame, and the independent variables were the skin-tone group, brightness, resolution, and motion. Resolution was converted from its label form (e.g., 288 x 192 pixels) into a single quantity: total pixels divided by 100,000. For the skin-tone groups, darker was coded as 1 and lighter as 0. Brightness and motion entered the model untransformed. Furthermore, frames from the same video were not treated as independent. If one video was dim and the tracker was unable to locate the hand for the majority of the video, the failure would appear as a significant number of consecutive missed frames. In reality, it would be just one video's worth of evidence rather than many separate observations. Treating all 1,149 frames as independent would make the sample size way larger than it is and give a false sense of accuracy. Therefore, the regression used standard errors clustered by video to account for correlation among frames within the same video. As a result, the confidence intervals were wider to show that the amount of independent evidence is provided by 50 videos rather than 1,149 individual frames.

4. Results

After averaging each video's detection rate within its group, the darker group had a rate of 71.9%, while the lighter group had 63.1% (Figure 2). The hypothesis was not just rejected, but, in fact, the results were the opposite.

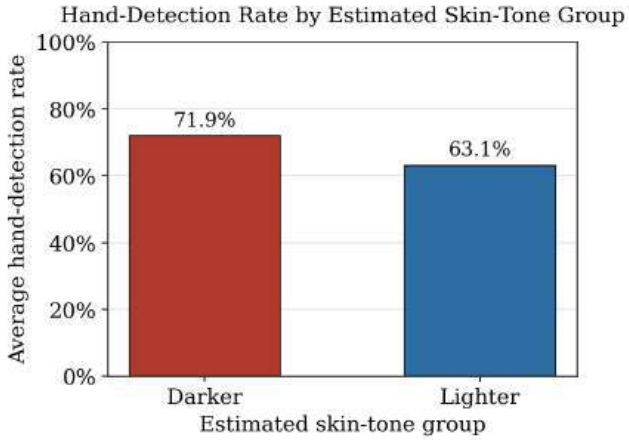


Figure 2. Hand-detection rate by estimated skin-tone group, first calculated per video and then averaged within each group

Cluster of Bright, Low-Resolution Videos

In Figure 3, each video is plotted as a singular dot, with detection rate on the vertical axis and average brightness on the horizontal axis. Most videos were spread along brightness values of roughly 25 and 140, with no clear distinction between the two skin-tone groups. However, at the top right of Figure 3, a tight cluster of 11 videos formed at brightness values above 190, and a nearly 100% detection rate. They also had the dataset's lowest resolution (288 x 192 pixels).

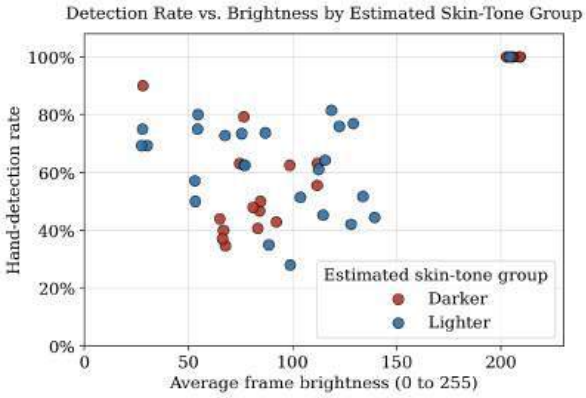


Figure 3. Per-video detection rate vs. average frame brightness, with the individual video dots colored by estimated skin-tone group (darker = red, lighter = blue). 11-video cluster visible at high detection rate and brightness above 190.

The cluster's high brightness and high detection follow the expected relationship. However, these are the lowest-resolution videos in the dataset, and that characteristic should hurt detection. Despite this, they detect almost perfectly, at nearly 100%. Every other video in the sample fell significantly below the cluster. Nothing about skin tone explains this, so to test whether that result depended on this cluster, an identical regression was estimated with the 11 videos excluded. The exclusion criterion was a resolution of 288 x 192 pixels and an average brightness above 190.

This first regression, the one before the exclusion, tested whether the darker group's higher detection rate remained after brightness, motion, and resolution were accounted for. Frames from videos in the darker group were still more likely to capture a hand, with a coefficient of +0.61 and a p-value of 0.020 (Table 1). Such a p-value reveals that if skin tone truly had no effect on detection, a coefficient this large would appear only about 2% of the time by chance, meaning the result is not likely to be an accident. Next, as expected for brighter and higher-resolution videos, they showed significantly higher detection while also having a p-value of less than 0.001. Furthermore, motion was barely significant ($p = 0.049$) and had a coefficient of +0.011.

Excluding the Cluster Removes the Apparent Effect

Excluding the 11 videos in the cluster left 15 darker-grouped videos and 24 lighter-grouped ones. With the cluster excluded, the skin-tone effect disappeared. The darker-skinned group's coefficient fell from +0.61 to +0.22, and its p-value rose tremendously from 0.020 to 0.442. Brightness ($p = 0.003$) and resolution ($p < 0.001$) stayed significant, similar to the previous model. Motion, which was borderline significant before, lost its significance entirely ($p = 0.219$).

Table 1. Two logistic regressions predicting per-frame hand detection: once for all 50 videos and once with the cluster of 11 videos excluded. Coefficients are log-odds estimates and standard errors clustered by video.

	Full data (50 videos, 1,099 frames)		**Cluster excluded (39 videos, 892 frames)**	
Independent Variables	***Coefficient***	***p***	***Coefficient***	***p***
Darker group	+0.607	0.020	+0.218	0.442
Brightness	+0.033	<0.001	+0.021	0.003
Resolution (pixels / 100,000)	+0.180	<0.001	+0.135	<0.001
Motion	+0.011	0.049	+0.013	0.219

5. Discussion

In both models run in this study, the two independent variables that mattered were brightness and resolution. Brighter video and more pixels meant more detected hands in both regressions. Estimated skin tone only seemed to matter when 11 unusual videos were included in the experiment; when they were excluded, the effect was gone. The cluster's almost perfect detection was extremely inconsistent with the rest of the videos in the sample. Furthermore, Roh

et al. report that MediaPipe misses over 50% of the hands in WLASL (323), again backing that the cluster was far from normal. Since the videos stood out so sharply, the regression was run again without them. Because the darker group's higher detection rate relied on this cluster rather than on an overall pattern from all 50 videos, this study found no evidence of a skin-tone effect.

If hand detection had failed because of skin tone, the solution would require slow, expensive changes to the model, including enhancing the algorithm, retraining on more diverse data, and testing the new model. Recording conditions are significantly easier to improve. Brightness can be increased by adding light fixtures, such as a lamp or a smartphone's flashlight. Resolution can be increased with a higher-quality camera, one that most modern smartphones already come with. These improvements are simple changes that the signers can make before recording.

By isolating the hand-tracking stage, these results open up important comparisons with existing research focused on later stages of sign-language recognition systems. Atwell et al. measured the final step and found that recognition accuracy was higher for lighter-skinned signers (5). Although this study did not directly evaluate recognition accuracy, it found no evidence that hand detection could be the reason for that bias towards lighter-skinned signers, as detection was similar across the two skin-tone groups once video conditions were accounted for.

As a clarification, at first glance, Figure 3 might seem to show the opposite of these results because the dots do not rise as brightness increases, making brightness seem like it does not impact detection at all. However, that is due to the effect of resolution—the brightest clips are also the lowest resolution ($r = -0.79$ across the 50 videos). The scatter plot cannot separate the two, so the regression shows the results more clearly: in Table 1, brightness and resolution are statistically significant and have positive coefficients, meaning that as they increase, so does detection.

Limitations

The most significant limitation is the skin-tone estimation. The estimator used can be biased towards lighter tones and confounded by lighting (Feng et al. 3). Although brightness can be isolated as much as possible to see its impact in the regression, the skin-tone estimation will ultimately be influenced slightly by the brightness. However, the main finding in this study—that brightness and resolution influenced detection significantly—stays solid because those two variables are measured directly from pixels, not determined by an estimator. Splitting the videos into two groups, lighter and darker, also has a limitation, as "darker" just means the darker side of the median, not a predetermined split.

Furthermore, the sample itself also has some limitations. Sample selection was not random, as the first 80 clips listed in the WLASL-100 index were attempted for use. Since 30 clips were deprecated or unusable, the 50 available were retained. Therefore, the results from these 50 clips cannot be generalized to all of WLASL.

Lastly, the measurements were deliberately simple, so each simplification has its drawbacks. This study measured only whether a hand was found at all, not accounting for the

accuracy of the hand keypoints placed. Many signs require two hands, and since it is a success (hand_present = 1) if just one hand is detected, it means this study just focused on locating one hand. Brightness was determined by taking in the entire frame, leaving a slight possibility for influence from backgrounds and clothing. Motion was calculated as the mean pixel difference between every third frame. Because higher-resolution frames contain more pixels, the same movement can produce a larger motion value in a higher-resolution video. Still, with a p-value of 0.049 in the initial regression and an even larger value of 0.219 in the second regression, motion went from barely significant to not at all, and it did not hurt the main finding. Other factors that were not measured, such as camera distance, could have influenced the results.

6. Conclusion

This study examined the first step of a sign-language recognition system—finding the signer's hands—and asked whether it fails more often due to skin tone or video conditions. Research on these systems has so far focused on the final step of the recognition pipeline, asking whether the output answer is correct. Analysis of this first step of the pipeline was important because if this first step fails, every subsequent step fails too. Based on the sample of 50 videos, this study did not find evidence that skin tone plays a role in hand detection; instead, it found evidence for brightness and resolution. Initially, the results communicated that darker videos had a higher detection rate, with hands in the darker group being detected in 71.9% of frames, compared to 63.1% in the lighter group. However, that was due to a cluster of 11 bright and low-resolution videos with high detection rates. Once those videos were excluded, the skin-tone effect disappeared while brightness and resolution remained significant; the higher the resolution or brightness was, the higher the detection rate.

This result matters in context because if the tracker had failed due to skin tone, the solution would require changes to the algorithm, training on a new dataset, and retesting. Because brightness and resolution follow a positive relationship with hand detection, the solution is easier. The model can stay unchanged, and signers only need a brighter setting and a higher-resolution camera.

Works Cited

Atwell, Katherine, et al. "Studying and Mitigating Biases in Sign Language Understanding Models." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 268-83.

Bragg, Danielle, et al. "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective." Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, Association for Computing Machinery, 2019, pp. 16-31.

Feng, Haiwen, et al. "Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation." Computer Vision – ECCV 2022, edited by Shai Avidan et al., Springer, 2022, pp. 72-90.

Li, Dongxu, et al. "Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, IEEE, 2020, pp. 1459-69.

Parte, Pablo, et al. "Auditing Demographic Bias in Facial Landmark Detection for Fair Human-Robot Interaction." arXiv, 10 June 2026, arxiv.org/abs/2604.06961.

Roh, Kyunggeun, et al. "Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition." Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, ELRA and ICCL, 2024, pp. 323-34.

Wilson, Benjamin, et al. "Predictive Inequity in Object Detection." arXiv, 21 Feb. 2019, arxiv.org/abs/1902.11097.

Middle East Underdeveloped? A long-duration Exploration of Ottoman Decline and the Limits of Modernisation By Mir Hamid

Abstract

Among the longest-lasting and largest empires in world history is the Ottoman Empire, which ruled large expanses of land stretching from Southeast Europe to Western Asia and parts of North Africa for about six hundred years. Despite its long existence and initial strength, however, the Ottoman Empire was unable to achieve industrialization on a scale witnessed in other parts of Western Europe. In the course of this study, therefore, the approach used is a longue durée one. It explores the way in which certain structural constraints in terms of institutional and socio-economic arrangements hindered any change over time within the empire. Instead of considering only the end stages of the empire's collapse, the process of industrialization and its potential role in uplifting the economy is analyzed, as is the way in which various obstacles – structural, political, religious, and demographic – hindered this process. The integration of the empire into the European capitalist system during the eighteenth and nineteenth centuries was characterized by rapid changes in world history as a result of the Industrial Revolution. Therefore, the modernization of the Ottoman Empire can be regarded as a complicated yet somewhat successful process that was bound to face certain limitations from external and internal perspectives. Through an examination of such limitations and their influence on the empire, it is possible to develop an approach to the problem of limitations of the industrialization and economic development processes in modern-day Middle Eastern countries.

Keywords: Ottoman Empire, Middle East, Industrial Revolution, Modernisation, Dependency.

Introduction

The Ottoman Empire stands as one of the most enduring and geographically expansive polities in world history, governing vast territories across Southeast Europe, Western Asia, and North Africa for over six centuries. At its height in the sixteenth and seventeenth centuries, it commanded strategic trade routes, administered diverse populations, and maintained sophisticated military and bureaucratic institutions. Yet, despite its longevity and early strength, the empire failed to undergo industrialisation on a scale comparable to that of Western Europe. This divergence contributed to widening economic and technological asymmetries that reshaped global power balances in the nineteenth century. Understanding why industrialisation did not transform the Ottoman economy is therefore central not only to Ottoman historiography but also to broader debates concerning development, modernisation, and the persistence of institutional constraints in the Middle East.

Despite the substantial body of scholarship that addresses Ottoman decline and economic integration, much of the literature examines the factors of decline in isolation or focuses narrowly on late-period crises. There remains a need for a more synthetic analysis evaluating how structural, political, religious, and demographic barriers collectively constrained

industrialisation as a potential mechanism of economic uplift. Rather than concentrating exclusively on terminal phases of imperial collapse, this study adopts a longue durée perspective, tracing how enduring institutional arrangements and socio-economic patterns limited transformation over time.

Furthermore, the relevance of this inquiry extends beyond the historical case into the contemporary Middle East. Three persistent domains of constraint emerge across both contexts. First, structural barriers, including institutional arrangements, property-right regimes, and state-market relations, influence investment, productivity, and innovation. Second, political constraints, such as governance instability, elite priorities, and reform capacity, shape economic trajectories. Third, social and demographic dynamics, encompassing education systems, labour force composition, and population pressures, affect technological adaptation and industrial development. By analysing how these factors operated within the Ottoman Empire, this study provides a framework for understanding why similar challenges continue to limit industrial and economic development in modern Middle Eastern states. Through this dual historical–comparative approach, the project aims to contribute to both Ottoman economic history and contemporary development studies. In doing so, it highlights how historically rooted constraints may persist, evolve, or reappear in new forms, shaping patterns of industrialisation and economic growth across time

Ottoman Decline and European Industrialisation

The Ottoman experience cannot be reduced to narratives of sudden decline or isolated military defeats, with many eagerly citing turning points such as Lepanto as pivotal moments of cataclysm that irrevocably set the Ottoman decline. Rather, the empire's economic trajectory reflects long-term structural conditions, political choices, social dynamics, and demographic pressures. Pasha Lotti, former Grand Vizier, having famously cited corruption as one of the greatest troubles of the empire, said, 'For officers of the state, corruption is a disease without remedy ... beware, beware of corruption; O God! save us from it.' (LEWIS, 1962)

A recurring theme in the historiography of Middle Eastern economic history is the region's perceived marginalisation within the global economy. As Samir Amin argued, "The development of some regions has historically required the underdevelopment of others" (Pamuk, 1984). This perspective positions the Middle East within an unequal world system shaped by external capitalist expansion. While such interpretations offer valuable insights, they risk overstating external determinism and underestimating the role of domestic institutions and political agency. The concept of dependency must therefore be problematised: to what extent were constraints imposed from outside, and to what extent did internal structures condition or reinforce economic limitations?

This tension is particularly evident in the Ottoman Empire's encounter with European industrialisation. The empire's increasing integration into European capitalism during the eighteenth and nineteenth centuries coincided with the Industrial Revolution, a transformative period that redefined global economic hierarchies. Expanding trade relations, capitulatory

privileges, and financial entanglements undeniably reshaped Ottoman production, fiscal stability, and market structures. However, external pressures alone cannot sufficiently explain the empire's limited industrial transformation. Structural rigidities, political fragmentation, and the configuration of religious, educational, and administrative institutions shaped the empire's adaptive capacity. Consequently, the Ottoman case presents a critical opportunity to analyse how internal and external dynamics interacted to influence economic outcomes.

The industrial limitations of the empire, although the empire had undergone a limited revolution, are argued to have been damaged by that of European industrial progression. As European goods were entering the Ottoman market in growing quantities after 1815 (Clark, 1974), these imports displaced traditional Ottoman products and put Ottoman specialised craftsmen out of work, and increased European influence over the Ottoman market. Further, such a pattern of imports also reduced the internal tax revenue of the state. Subsequently, the Ottoman state did make a serious attempt to respond. It has been argued that Ottoman industrialisation efforts, especially in the 1840s, have been underestimated (Clark, 1974). Earlier reforms under Selim III and Mahmud II introduced European machinery and factory methods, mainly for military production such as cannon, rifles, cloth, paper, boots, and fezzes. However, these early efforts were limited and focused only on final manufacturing rather than the wider industrial system(Clark, 1974). The 1840s were different, as during the early Tanzimat, the Ottoman state launched its largest and most ambitious industrial programme, centred around Istanbul, Izmit, Hereke, and Bursa. This included foundries, cloth mills, silk works, machine shops, schools, model farms, mines, and efforts to secure domestic raw materials like iron, coal, wool, cotton, and silk (Dawn, 1989). Clark argues this as the Ottoman Empire's prime hope for a true industrial revolution. However, the empire had already been deeply inadequate in its position to capitalise on such a revolution, with outdated foundational laws still having never been modernised; for example, cannons were never given an official regulation size (Clark, 1974).

Ottoman Industry and the Question of Innovation

Bernard Shaw argues that the Ottomans made no meaningful attempt to modernise the systems through which reform could be facilitated, having been so severely inhibited that such reforms could not take place. The modernisation in question was far too little and far too late to be seen as a real attempt (Bernard, 1962). With the Tanzimat reforms having taken place long after the industrialisation of Europe - as such the difference in technological advancement between Europe and the Empire was already increasing exponentially, so the gap was seen as having been far too large. The Empire had also failed to match the same level of development as in Europe, having failed to consistently match the same technological developments. Lastly, whatever reform did happen, it was always reactive and never proactive. (Bernard, 1962)

The Ottomans were never able to modernise in the same systemic capacity as their European rivals and counterparts. The Ottoman pattern of reactivity can be identified as far back in Ottoman history as 1571, when, after the crushing defeat at Lepanto, the Ottoman fleet was completely rebuilt in the same style and type of barge. Rather than innovating and changing the

primary vessel through which they conducted naval warfare, they opted to simply rebuild the same type of vessel. Even though the new vessel for naval conduct was now the more versatile 'barge', which relied on sails to travel, allowing for greater movement distance and the ability to house more weaponry and soldiers.

This is but one simple example of a lack of proactivity and ingenuity earlier in the Ottoman Empire. Modernisation remained on the surface level, not changing society or the economy deeply. While the Tanzimat focused on law codes, administration and military organisation, the reforms touched every aspect of Ottoman life. New secular courts were established to operate alongside religious ones, and legal codes inspired by European models were introduced to regulate commerce, crime, and civil affairs. The government sought to standardise taxation, curb corruption, and bring provincial authorities under tighter central control. The military, long a symbol of Ottoman weakness, was reorganised with modern training and conscription practices. Education, too, became a priority, with new secular schools and academies designed to produce a professional administrative class (Endelman, 2018). However, these were mostly reactive reforms as highlighted by Shaw. It can be argued that the modernisation was not as effective as in Europe, but it was still meaningful and facilitated positive change and advancement. While this reform was purely ‘top down’ (Bernard, 1962), having been driven from the top by the elite bureaucracy, it still led to change, and the change was further limited by internal blockages such as traditionalists and other conservative and religious groups, for instance, the ullema. It was simply an incomplete transformation (Endelman, 2018).

Shaw's theory, which had emphasised the decline, stagnation and failure of the Ottoman state to make real meaningful efforts to modernise, fails to account for the external pressures that would have impaired the Empire's ability to modernise in the same depth and to the same extent as Shaw believed they should. With Shaw having essentially argued that they did not put as much effort into modernising as they should have, however, this interpretation has been increasingly anachronistic, with the assumption that Ottoman modernisation was inherently unsuccessful being far more unpopular. Many of the problems faced by the Ottomans during this period of reformation were not as strongly a result of their own indolence and incompetence as previously suggested. Instead, scholars like Pamuk argue that the empire underwent significant institutional, economic, and political transformation, and that its limitations were shaped less by internal incapacity than by global economic pressures and historiographical bias. Consequently, rather than viewing Ottoman modernisation as a failed project, it is more accurate to interpret it as a complex and partially successful process constrained by external forces and misrepresented by earlier historical narratives.

Dependency Theory and the Consequences of Ottoman Decline

The first and most central weakness in Shaw's argument is its insistence upon placing the blame for the reformations' shortcomings on the Ottoman state itself, and not accounting for the several external constraints under which the empire was forced to operate. The works of Pamuk

are critical in understanding Ottoman limitations through 'dependency theory'. Pamuk argues that the Ottoman Empire's economic trajectory was heavily shaped by its integration into the expanding global capitalist system, which placed it in a structurally disadvantaged position, with examples of such a position being the exponential inflation, the empire suffered of an increase in the circulation of silver in the global market; having come from the new world, which caused rampant inflation in the Ottoman Lira which was made of silver. Fluctuations in international trade, unequal terms of exchange, and increasing European economic dominance also limited the empire's capacity to industrialise on its own terms. Ottoman modernisation should be viewed as the Ottoman economy being constrained rather than underdeveloped or unsuccessful.

The historical constraints that obstructed Ottoman industrialisation have not disappeared with the empire itself; rather, they have survived in altered forms and continue to shape the Middle Eastern landscape today. The region's geopolitical fragility, uneven economic development, and persistent dependence on external powers can all be traced, at least in part, to the institutional and structural weaknesses that became entrenched during the final centuries of Ottoman rule. The empire and its subsequent decline had led to a collapse in structural stability within the region, and the power vacuum that had replaced the structure they had built had a lasting impact. As the influence of foreign parties could be more heavily exerted over these regions, as detailed earlier with Pamuk's dependency theory. Structurally, one of the most enduring legacies is the weakness of autonomous industrial and productive development. Much of the Ottoman economy became increasingly tied to European markets as a supplier of raw materials and a consumer of manufactured goods, a pattern which survived the empire's collapse and continued under colonial and postcolonial rule. Many modern Middle Eastern economies still remain heavily dependent on a narrow set of exports, especially oil and gas, while lacking diversified industrial bases capable of producing sustained, broad-based development. This reflects a continuation of the same imbalance that Pamuk identifies in the Ottoman case: integration into global capitalism under unequal terms. Rather than fostering self-sufficient economic growth, external economic linkages have often reinforced dependency, leaving states vulnerable to global price shifts, foreign capital, and international political pressure.

Politically, the Ottoman legacy of uneven centralisation and reform from above also remains highly significant. As seen in the Tanzimat period, reform was frequently state, and although the Empire itself was not fully to blame for the inadequacies in such a reform, these inadequacies were still present; this method of reform from the top down coupled with the lack of long term structure within the reform lead to the foundations being set within the middle east of a reform that was elite driven, and limited by entrenched interests, rather than rooted in broad institutional transformation. This pattern has echoed across many modern Middle Eastern states, where development projects have often depended on authoritarian central governments attempting to modernise without fully restructuring political participation, social representation, or administrative accountability. The result has frequently been a shallow form of modernisation such as infrastructure without institutional depth, legal reform without social transformation (in that a far more religiously driven legal system had still not adapted to an increasingly nuanced

and financially dominated world), and state-building without durable legitimacy. Such conditions have contributed to repeated instability, weak governance, and recurring crises of state authority across the region.

There are also further demographic and social dimensions that are equally important. In the Ottoman Empire, educational inequality, limited technical training, and the fragmented social composition of the population reduced the empire's capacity to adapt to industrial modernity. Similar issues remain visible today in many Middle Eastern states, where education systems do not always align with market needs and where youth unemployment remains high, with a UN survey conducted in 2022 having put the unemployment rate of the Arab regions at approximately 12%, which was one of the highest in the world, given that the UK unemployment rate was around 4.9%. Further, these regions are where sectarian, ethnic, and regional divisions continue to complicate national cohesion (UN News, 2022), with huge majorities of people having been underrepresented in the creation of borders, which leads to an abundance of civil unrest. A prime example of this is the Kurds and their subsequent drive for independence from Iran and Turkey. These social constraints weaken the long-term development of human capital, which is essential for industrialisation, innovation, and sustainable growth. Just as the Ottoman state struggled to translate reform into systemic economic change, many modern states in the region struggled to convert population growth and urbanisation into productive industrial capacity.

The trajectory of Ottoman institutional development from its successful patrimonial foundations through the structural transformations of the late sixteenth and seventeenth centuries to the reactive modernisation of the Tanzimat period reveals a pattern of continuity that defies the sharp periodisation of the conventional decline narrative. The institutional weaknesses that ultimately constrained Ottoman industrialisation, seen in the erosion of the devshirme, the corruption within the political system, the breakdown of the timar system, the withdrawal of sultanic personal governance and the growing influence of entrenched interest groups and factions, were not the sudden products of post-Suleimanic decay but were embedded in the structural systems of the empire across centuries.

Conclusion

These problems did not disappear when the Empire collapsed following the conclusion of the First World War in 1922. Across the twentieth and twenty-first centuries, many Middle Eastern states have continued to face similar challenges: reliance on commodity exports, high youth unemployment, weak public trust in government, and vulnerability to outside powers. This suggests that the empire's institutional weaknesses, made worse by economic dependency, were not simply historical failures of the Ottoman period. They were inherited by the states that followed and still affect the region's development today. In this sense, the Ottoman Empire is not just a case study from the past; it helps explain why institutional weakness and economic dependency remain so deeply connected in the modern Middle East.

Works Cited

Arab region registers world's highest unemployment rate, UN survey finds | UN News. (2022, December 30). News.un.org. https://news.un.org/en/story/2022/12/1132112

ATTİLA, A. E. (2009). Agrarian Relations, Property and Law: An Analysis of the Land Code of 1858 in the Ottoman Empire. Middle Eastern Studies, 45(6), 935–951. JSTOR. https://doi.org/10.2307/40647178

Clark, E. C. (1974). The Ottoman Industrial Revolution. International Journal of Middle East Studies, 5(1), 65–76. JSTOR. https://doi.org/10.2307/162344

Dawn, C. E. (1989). The Ottoman Empire and European Capitalism, 1820–1913: Trade, Investment and Production. By Şevket Pamuk · New York: Cambridge University Press, 1987. xii + 278 pp. Maps, charts, tables, appendixes, notes, bibliography, and index. $42.00. Business History Review, 63(1), 232–234. https://doi.org/10.2307/3115454

Deringil, S. (2003). "They Live in a State of Nomadism and Savagery": The Late Ottoman Empire and the Post-Colonial Debate. Comparative Studies in Society and History, 45(2), 311–342. https://doi.org/24/05/2026

Dietz, J. L., Cardoso, F. H., Faletto, E., & Frank, A. G. (1980). Dependency Theory: A Review Article. Journal of Economic Issues, 14(3), 751–758. JSTOR. https://doi.org/10.2307/4224952

Endelman, J. (2018). In the Shadow of Empire: States in an Ottoman System. Social Science History, 42(4), 811–834. https://doi.org/10.1017/ssh.2018.3

Finkel, C. (2007). Osman's Dream: the Story of the Ottoman Empire, 1300-1923. Basic Books.

LEWIS, B. (1962). OTTOMAN OBSERVERS OF OTTOMAN DECLINE. Islamic Studies, 1(1), 71–87. JSTOR. https://doi.org/10.2307/20832621

Pamuk, Ş. (1984). The Ottoman Empire in the "Great Depression" of 18731896. The Journal of Economic History, 44(1), 107–118. JSTOR. https://doi.org/10.2307/2120558

Quataert, D. (2000). Consumption studies and the history of the Ottoman Empire, 1550-1922 : an introduction. State University Of New York Press.

Nixon's Obsession With Image Control: The Watergate Scandal By Victoria Altirs

"That's my boy! That's my actor," Richard Nixon's former acting coach Albert Upton exclaimed as Nixon went on national live television and cried after his name was cleared as a result of his Checkers Speech (Thomas 4). In college, Nixon acted in several plays, learning how to grow his emotional range. Later, during his presidency, Nixon used these acting skills not only to try to cover for his mistakes but to create an entirely new persona that was only shown to the public. Although he projected a cool and calm image, Nixon was driven by fear, believing everyone around him was out to get him (Thomas xiii). He preferred to be alone and thrived when it was only him and his ideas (Reeves 12). His new public persona was not driven by fear like the real Nixon was; this persona was strong, powerful, and composed. However, after the Watergate scandal, Nixon's facade started to slowly crumble under the weight of fear. On June 17, 1972, five men hired from Nixon's election committee were caught by police breaking into the Democratic Party headquarters at the Watergate with electronic bugging devices (Emery xii). This affair was then dubbed the Watergate scandal and went on to define Nixon's presidency. In an attempt to curate a perfect public image, Nixon continuously manipulated information and antagonized the press, leading not only to his downfall but to a decline in public trust in government.

Nixon's early campaigns revealed how deeply he relied on deceptive tactics and media control to protect his image and win elections. In 1946, while running for California's 12th Congressional District seat, Nixon showed the public a pamphlet containing a series of endorsement recommendations from the National Citizens Political Action Committee (NCPAC). This falsely linked his opponent, Jerry Voorhis, to the CIO Political Action Committee (PAC) (Stone and Colapietro 75). Due to the CIO PAC's associations with communism, many voters perceived the CIO PAC's alleged endorsement of Voorhis negatively, which aided Nixon in winning the election. In 1947, Nixon acknowledged his deception in misleading the public, admitting to Voorhis's advisor, "Of course, I know Jerry Voorhis wasn't a communist... I suppose there was scarcely ever a man with higher ideals than Jerry Voorhis, or better motivated...but...I had to win. That's the thing you don't understand. The important thing is to win" (Stone and Colapietro 75). Winning was the only thing that mattered to Nixon. To win a seat in the House of Representatives would purely increase Nixon's power, which would allow him to manipulate people on a grander scale. As his status grew, his influential outreach would be boosted. Nixon's willingness and ability to compromise ethics for political gain showed his obsession with image and success.

Nixon used manipulation to escape scandals. Just weeks before the 1952 presidential election, where he was running for vice president on a ticket with Dwight D. Eisenhower, Nixon was accused of misusing a secret political fund of about $18,000 that he had received from his supporters by spending it on personal campaign reimbursements (Stone and Colapietro 101-105). This revealed not only his willingness to accept money from shady sources but also his readiness

to lie to the public about his financial priorities. Nixon denied these allegations in his television address known as the Checkers speech. In the speech, to manipulate his audience and evoke sympathy, he shared a story:

> We did get something, a gift, after the election…the day before we left on this campaign trip, we got a message from Union Station in Baltimore, saying they had a package for us. We went down to get it. You know what it was? It was a little cocker spaniel dog, in a crate …our little girl, Tricia, the six-year-old, named it Checkers. And you know, the kids, like all kids, love the dog, and I just want to say this, right now, that regardless of what they say about it, we're gonna keep it. (Nixon)

Nixon's sympathetic story and direct eye contact proved to be effective, as when Nixon returned to his hotel, he was met with a crowd of supporters who had listened to his speech over the public address system. When Congressman George H. Bender asked the crowd about their support for Nixon, they responded by chanting, "We want Nixon!" (Stone and Colapietro 107). The Checkers Speech revealed to Nixon that he could use emotional manipulation to evade public blame or suspicions. By directly appealing to voters on television with evocative, unthreatening language like "our little girl" and the promise that the family would "keep" the dog, Nixon attempted to convince the public of his good character. Even Lyndon B. Johnson, a fierce political competitor of Nixon, admitted that Nixon was able to use television to reinvent himself as a more likable man, describing Nixon as "the only politician in the country who was more polarizing" than himself (Stone and Colapietro 237).

The fact that the Eisenhower-Nixon ticket then went on to sweep thirty-nine states and win the popular vote by six million was proof that Nixon's Checkers speech saved his political career (Stone and Colapietro 103). However, it led many intellectuals to ridicule him as overdone and shallow (Farrell 208).

Nixon continued his pattern of manipulation by playing the victim in situations he created. Ten years later, in the 1962 press conference after Nixon lost the governorship of California, he again began playing the victim and pitying himself in order to emotionally manipulate his public perception, saying of his opponent, Pat Brown: "I wish him well. I wish him well, not only from the personal standpoint, because there were never on my part any personal considerations. I believe Governor Brown has a heart, even though he believes I do not. I believe he is a good American, even though he feels I am not" (Nixon). Nixon's obsession with curating a perfect public image is evident in this quote, painting himself as an empathetic and supportive competitor. During the same 1962 press conference, Nixon went on to say,

> But as I leave you, I want you to know- just think how much you're going to be missing. You won't have Nixon to kick around any more, because, gentlemen, this is my last press conference, and it will be one in which I have welcomed the

> opportunity to test wits with you. I have always respected you. I have sometimes disagreed with you. (Nixon)

By describing himself as someone the press could “kick around,” Nixon both positioned himself as a mistreated victim and set the tone for a hostile relationship with the press.

After realizing the significance of polls in elections, the Nixon administration pressured pollsters and manipulated results to sway public opinion, with the intention of making Nixon appear more favorable. In 1969 and the early 1970s, Nixon and his top aides used strong-arm tactics against pollster Louis Harris. Nixon mistrusted Harris because of Harris’s affiliation with the Democratic Party, as Harris had worked as a pollster and consultant for John F. Kennedy during the 1960 presidential campaign. After a debate in 1960 with Kennedy, the Gallup Poll concluded that 43% of the public thought Kennedy won the debate, 29% said it was a toss-up, and 23% handed it to Nixon (Farrell 285). The debate was a big factor in Kennedy winning the election. This led Nixon to realize the support he could get when the polls were on his side. However, Nixon was suspicious of his low poll numbers and viewed Harris as an enemy working for "the other side," a phrase coined by Harry Dent, a political advisor for the Nixon administration, while referring to the Kennedy Party (Jacobs and Shapiro). Since Louis Harris and George Gallup were leaders in polling, it drew Nixon's attention and made them targets for his administration.

Due to Nixon’s insecurities, he wanted to use the polls to help control the narrative surrounding him. This was particularly important to Nixon. According to Nixon’s speechwriter Raymond K. Price in his late 1967 campaign strategy memo, Nixon’s biggest challenge was overcoming the belief that he could not win. An idea that he seemed to internalize regardless of poll results (Stone and Colapietro 311). Nixon was self-conscious and self-obsessed, so to make himself feel better, he needed a sure way to appear as the favorite candidate to the public. Nixon and his administration viewed polling not as a tool for gauging public opinion, as it was intended for, but as an outlet where controlling the results of the polls meant controlling the narrative.

Nixon’s desperation to alter the polls in order to control his image led to his forced control over Harris and Gallup. Nixon’s eagerness was revealed in his chief of staff, Harry Robbin (H.R.) Haldeman’s diaries, where he wrote that Nixon believed Harris and Gallup’s poll results "directly affect our ability to govern" due to their influence on Congress members and foreign leaders. According to Haldeman's diaries, Nixon initially aimed to "zing" Harris and halt his use for departmental polls, and the administration sometimes pressured Gallup by contacting newspapers to suppress polls. However, finding that the polls were generally politically sympathetic to Republicans, the Nixon administration sought to reshape polling results by manipulating the pollsters themselves, often appealing to pollsters' patriotism by linking cooperation to national duty (Jacobs and Shapiro 521). In fact, both Louis Harris and George Gallup admitted in an interview with reporter Lawrence Jacobs that they cooperated with every president who approached them, saying they considered it “a public service to the country”

(Jacobs and Shapiro 522). This was essentially guilting pollsters into bending to the will of the Nixon administration, whom they could then manipulate further for their own benefit.

In 1969, on top of pressuring pollsters, Nixon continued his attempts to control his portrayal in the media. He was convinced that most journalists had a "strong negative attitude" toward him and wanted to sabotage his government. Nixon's mistrust of the press was demonstrated by derogatory comments he made about reporters he did not like to Haldeman, like calling CBS correspondent Marvin Kalb a Communist (Patch 555). When Nixon was finally elected, his opinion of the press only worsened during the Vietnam War. Many news stories led Nixon to accuse journalists of failing to give him credit for reversing the war's trajectory, making him believe that they were rooting for the enemy, Vietnam, to see him fail (Pach 555). On June 5, 1969, after Nixon ordered 25,000 American troops home because of the growing strength of the South Vietnamese army, CBS News broadcasted stories showing South Vietnamese troops panicking in combat, and aired an interview with U.S. advisers where they doubted the South Vietnamese troops' ability to defend the country. This directly contradicted Nixon's claims that South Vietnam was ready to take over the war effort. Then, when Nixon's critics claimed he was either extending the war in Cambodia and Laos or removing American forces too slowly, Nixon and his associates responded by accusing the media, particularly television networks, of broadcasting false information. Nixon even went so far as to claim that the administration's worst enemy seemed to be the press during the war in Vietnam (Pach 555-556).

In addition to controlling the press, Nixon made efforts to protect his own image by appointing trusted friends and accomplices to positions of political power. The administration had realized the weight of the press and needed another way to control the polls. This was true in late 1969, when the White House recognized that Dan Lufkin, a friend close to Nixon, held "a controlling interest in the Harris Poll." Harry Dent had recruited Peter Flanigan, a consultant in the Nixon administration, to persuade Lufkin to influence the firm. Flanigan reported that Lufkin promised to "keep [Harris] honest," and the White House also sought to publicly challenge Harris (Jacobs and Shapiro 522). The White House followed through on its threat in June 1970, when Haldeman instructed a senior White House aide, Charles Colson, to investigate pollsters in the United States under the pretext of inaccurate polling in a recent British election. More specifically, Colson was directed to "zero in on Louis Harris' polling organization" and to make the project a top priority, one that was well-publicized on the news but would not be directly linked to the White House (Jacobs and Shapiro 522). This investigation pressured Harris to fall back in line with the Nixon Administration's demands, delivering data that would be favorable.

Since it was clear that polls played a crucial role in the election process and in conveying the opinion of the public, Nixon's team shifted from punishment to strategic bribery. This was because having a high support shown through the polls would bolster Nixon's sense of security and possibly secure the vote for some members of the public. By 1971, the administration was providing contracts as incentives, as Lufkin suggested to Flanigan: "It will be easier to ensure Republicans get a fair shake if Harris occasionally receives a Republican contract" (Jacobs and Shapiro 522). Shortly afterwards, Harris collaborated with Colson on a survey to collect public

opinions, receiving $39,000 in government funds as payment. These tactics demonstrate how far Nixon was willing to go to manipulate information, all to maintain the image of a politician who was more popular than the facts allowed.

When Nixon's political standing changed as a result of feedback from the Vietnam War, he demonstrated resentment towards not only pollsters but also the press. In 1971, when Nixon fell behind Senator Edmund Muskie, the Democratic presidential frontrunner, in the polls, he attempted to regain control by conducting press and TV interviews. However, it was clear he lacked media support. Only a few friendly journalists were willing to interview him. There was a history behind the tension in Nixon's relationship with the press, which had been marked by mutual disdain (Emery 37). With his constantly changing opinion on the press and mutual hostility in his early political years, Nixon only increased tensions and fueled the distrust that would later explode during Watergate.

Nixon and his staff repeatedly lied and obstructed the Watergate investigation to hide the truth, ultimately contributing to the severity of the scandal. When promoting the investigation was not enough, Nixon and his administration turned to tactics such as stonewalling, lying, and cover-ups to hide their role in the scandal. In June 1972, Nixon and White House Counsel John Dean used hush money to keep those involved in the cover-up silent (Basinger and Rottinghaus 193). Nixon was willing to go to any length to keep the scandal and his involvement hidden from the public. Seeing the payoff as a quick fix, Haldeman assured Nixon that there was no issue. In his memoirs, Nixon ironically noted, "It seemed that our intervention had worked easily. As far as I was concerned, this was the end of our worries about Watergate" (Emery 193). Nixon was confident in his manipulative tactics and prioritized his image over legal or ethical considerations. A number of people in Nixon's administration who were facing indictment, like G. Gordon Liddy, said they were coerced into remaining silent and entering guilty pleas. Nixon further impeded the investigation by dismissing Special Prosecutor Archibald Cox and pressuring White House employees to block the congressional investigation by refusing access to Oval Office audio recordings, alleging presidential privilege (Basinger and Rottinghaus 231-232). Firing Cox suggested that Nixon was abusing his powers as president and trying to stop Cox from getting close to the truth. By refusing access to recordings that would expose him and halting Cox's investigation, Nixon continued his efforts to hide the truth of his wrongdoing.

Building on the pattern of obstruction, Nixon contributed to his own downfall by escalating from lying to fabricating evidence. When talking with his attorneys, Joseph Fred Buzhardt Jr and Leonard Garment, Nixon said, "Why can't we create a new Dictabelt?" He suggested dictating a new recording from handwritten notes and submitting it to the court as the original. Buzhardt stated that this attempt to fabricate evidence was obstructing justice. Garment emphasized the gravity of the situation, noting that the President's willingness to manufacture evidence showed his desperation and blatant lack of respect for the law. Buzhardt's testimony in 1973 would confirm that the President had always intended to cover up and fabricate the truth (Woodward and Bernstein 27).

Continuing to go to great lengths to maintain his public image, Nixon's desperation to manage his reputation led him to project false public support of the investigation. The strategy constructed by Nixon's team was that by publicly supporting the congressional investigation, Nixon would not appear guilty, and it would prevent investigators from turning their attention onto him. In a press brief on March 21, 1973, Nixon stated that anyone in the executive branch who was indicted would be suspended, and those convicted would be discharged. Although this was obvious, he added a statement he believed improved the White House's image, despite its minimal public impact. He emphasized that no individual in a significant administrative role should receive immunity from prosecution (Emery 338). Days later, on April 22nd, 1973, *The New York Times* published an article about the affair and Nixon's apparent act of transparency in a scandal that he had created:

> High government officials found themselves scrambling to establish their innocence, or relative innocence, of the damning scandal. The President himself last week took a new line on the affair. No one in 'a position of major importance in the Administration,' he said, would be immune from criminal prosecution. The scramble for position immediately turned more public and frantic. ("Watergate")

By promising to prosecute any wrongdoers, Nixon projected an authoritative sense of control and an illusion of fair punishment, which hid his underlying motive of protecting himself.

Nixon released a statement full of lies in another attempt to protect his public image. On May 22, 1973, the President issued his second major statement within a month, declaring that he would not resign. However, this statement, which Haldeman later claimed was intended to provide a complete account, contained several lies, and it would serve as a benchmark against which the following news was measured. Six of Nixon's first seven points, which aimed to rebut or preempt allegations against him, were later proven to be false after he resigned. One of the most notable points was this: "At no time did I attempt or did I authorize others to attempt to implicate the CIA in the Watergate matter" (Emery 360). These points raised the suspicion of prosecutors Earl Judah Silbert and Seymour Glanzer, which then led to their exposure. This revealed that Nixon had deliberately lied to the whole nation; his primary motivation was to maintain his public image of moral and political integrity.

Despite Nixon's best efforts to paint himself in a positive light, relentless media coverage eventually exposed his lies and shattered his image, ultimately leading to his political downfall. On July 13, 1973, Alexander Butterfield, the deputy assistant to Nixon, confirmed former White House counsel John Dean's suspicions that Nixon had requested and installed a secret recording system. In a closed-door meeting with the Senate Watergate Committee, he said, "Yes, there was a taping system that recorded all presidential conversations" (A&E). Due to the fact that Dean testified to this during the Senate Watergate Committee's investigation, and the high media coverage following the investigation, Butterfield was forced to publicly reveal the existence of the system just three days later during his live television testimony. Although Nixon's close aides

confirmed the fact that there were tapes recording all of Nixon's conversations, Nixon still attempted to hide the evidence of his crimes. Citing presidential privilege, he refused to turn over his secret audio recordings to Special Prosecutor Archibald Cox and the Senate Watergate Committee on July 23, 1973 (Woodward and Bernstein 309). Nixon's rejection sparked a legal dispute that ended a year later when the U.S. Supreme Court unanimously decided that Nixon had to turn over the tapes (A&E). Nixon's initial refusal to hand over the tapes, despite Dean's confirmation of their existence, showed how desperate Nixon was to influence public perception right up until his crimes were revealed. As Basinger and Rottinghaus acknowledge in "Skeletons in White House Closets: A Discussion of Modern Presidential Scandals," presidents have the chance to actively collaborate with or work against the media (Basinger and Rottinghaus 230-231). Nixon's decision to "work against the media" was a direct reflection of his stubborn desire to control the public's narrative around him, a plan that backfired and led to his resignation.

As the Watergate story gained traction, Nixon's efforts to silence the press intensified because he knew the true story would reveal his crimes. The President was aggressively tackling news leaks about Watergate. According to a source in the Executive Branch dubbed "Deep Throat" by reporter Bob Woodward, Nixon instructed officials to "go to any length to stop them" (Bernstein and Woodward 269). He was serious about this, with internal investigations underway and plans to involve the courts. In a secret in-person meeting, Deep Throat revealed more information, saying Nixon was worked up, shouting in meetings, "We can't have it, and we're going to stop it, I don't care how much it costs" (Bernstein and Woodward 269). He believed the media had gone too far and needed to be reined in. To him, press coverage was not just criticism; it was a personal attack on him and his administration. Determined to address the issue, he saw it as crucial to the integrity of government and loyalty. He viewed the press as antagonistic and considered those who spoke with them to be enemies from within. Meanwhile, the press was becoming increasingly suspicious of the administration's honesty. In an ABC evening newscast, correspondent Don Farmer remarked, "We've been lied to so many times that you suspect no one tells the truth" (Pach 562). Similarly, CBS anchor Walter Cronkite accused the administration of orchestrating "a grand conspiracy to undermine the credibility of the press" (Pach 562). These comments reflected a turning point. As Nixon's need for control grew, the press's suspicion grew.

Due to the reveal of the taping system, Nixon's cover-up attempts ultimately fell apart. From February 16, 1971, to July 18, 1973, Nixon had recorded around 3,700 hours of meetings and phone conversations in the Oval Office, Cabinet Room, and Camp David (A&E). These hours of recordings were key to the trial and key to cracking Nixon's facade. In the transcript from one of the tapes presented to the jury during Nixon's trial, a recording of a meeting between H.R. Haldeman, John Dean, and Nixon, Nixon can be heard discussing the media's coverage of the scandal:

> DEAN: Uh, I think we can say "Well" at this point. The, uh, the press is playing it just as we expect.

HALDEMAN: Whitewash?
DEAN: No, not yet; the, the story right now –
PRESIDENT: It's a big story.
(Dean et al.)

This revealed that Nixon did indeed know what was going on regarding the scandal. As the president, Nixon was responsible for his staff, so his participation in a conversation about creating a "whitewash" demonstrates his intention to manipulate the story. Other conversations about hush money and knowledge about cover-ups recorded on the tapes revealed that the President had lied to the nation, his aides, and his lawyers for over two years. Although the tapes might not have been enough to prove legal guilt, they would have likely aided in the impeachment and conviction of Nixon (Woodward and Bernstein 309). Evidently, the tapes threatened Nixon's carefully crafted persona: they would reveal the insecure, fearful man he was and destroy his public persona.

Even when Nixon knew the end of his presidency was near, he still attempted to maintain a positive public image by taking responsibility for his actions. On November 17, 1973, at a news conference with the Associated Press Managing Editors Association, Nixon admitted to his mistakes when asked by reporter Mr. Quinn if the US could keep a republic after Watergate. "I recognize that because of mistakes that were made, and I must take responsibility for those mistakes, whether in the campaign or during the course of an Administration, that there are those who wonder whether this Republic can survive" (*The New York Times*). Nixon's statement showed he knew Watergate had badly damaged public trust, but by blurring the truth with misleading language, he avoided taking full responsibility while appearing as though he did.

In 1974, after Nixon was subpoenaed for the tapes, he attempted to control the narrative again by turning over his own transcript of the tapes, showing his continued attempts to uphold his public image. This was a way for Nixon to still hold control while claiming he was honoring the subpoena. The transcript revealed a conversation between Nixon and the White House Counsel attorney, John Dean, during which Nixon said, "'Tell the truth. That is the thing I have told everybody around here", and his famous line of foreshadowing, "I made it clear there was to be no cover-up" (Nixon 267). By publicly declaring his support for the investigation, Nixon masked his actual role in the scandal. His persona in the press drew few parallels to his actual character. On another of his transcripts, he turned in, he said,

> On April 15, I [Nixon] learned that Gordon Liddy's refusal to cooperate was hindering the case. I called [the Assistant Attorney General] and instructed him to inform Liddy, and everyone involved, that, quoting the tape, 'As far as the President is concerned, everyone in this case is to talk and tell the truth.' I offered to meet with Liddy's lawyer if necessary to stress the importance of his cooperation. (Nixon 266).

He also often ended his statements by saying, "I condemn any attempts to cover up, no matter who is involved," as he did on April 17, 1973, during a press conference, to reinstate and drill into the public's mind that he was fighting the fight for them (Emery 338).

Even though the tapes were subpoenaed, manipulation behind the scenes by the Nixon administration still occurred, shattering the public's confidence in Nixon. During the Watergate trials, the investigation of the U.S. government that was conducted due to the Watergate scandal, prosecutors noticed an eighteen and a half minute gap in the tape recording of a meeting held between Nixon and Haldeman on June 20, 1972 (*Hoover Institute*). This was three days after the break-in. Nixon's personal secretary, Rose Mary Woods, claimed that she had accidentally erased part of the tape when she was transcribing it because she paused to answer the phone. Since Woods was very loyal to Nixon and had worked for him for over twenty years, prosecutors were skeptical (*Hoover Institute*). They became even more skeptical when they had her reenact the scene and saw the absurd position she was in. This came to be known in the press as the "Rose Mary Stretch" because of how hard it was to believe and how ridiculous the explanation was; see fig. 1 for press imagery.

Fig. 1. Hoover Institution Library & Archives, "Un-Presidented Watergate and Power in America," Hoover Institution Library & Archives, https://www.hoover.org/library-archives/histories/un-presidented.

According to a *New York Times* article published on June, 5, 1974, a court-ordered investigation showed that a panel of technical experts concluded that "the only completely plausible explanation for the 18½- minute gap…is that the gap was caused by pushing the keys of a normally operated tape recorder at least five times" (Oelsner 89). Ultimately, the Judge in the case, Judge John J. Sirica, suggested that there had been "unlawful tampering" with the tape (Oelsner 89). The scandal of the gap wrecked any of the little credibility Nixon had left, as it had

become increasingly clear how desperate he was to conceal the truth and how little respect he had for the law.

Nixon rode the line of defending himself while also appealing to the public until the end of his presidency. In his last speech as president on August 9, 1974, Nixon said:

> ...As I pointed out last night, sure, we have done some things wrong in this Administration, and the top man always takes the responsibility, and I have never ducked it. But I want to say one thing: We can be proud of it - 5 ½ years. No man or no woman came into this Administration and left it with more of this world's goods than when he came in. No man or no woman ever profited at the public expense or the public till. That tells something about you… Mistakes, yes. But for personal gain, never. You did what you believed in. Sometimes right, sometimes wrong. (Nixon 271).

He admitted that Watergate was a mistake, but he avoided the narrative that it was for selfish intent. Nixon had painted himself as a good public servant despite repeatedly trying to rise above the law and manipulate the public and those around him.

After Watergate, the public lost trust in the government because of what Nixon put them through. The loss of trust stemmed directly from Nixon's manipulation and dishonesty, revealed during the scandal. As a result, Nixon no longer deserved the confidence that many Americans, especially those who supported him in 1972, had in him after Watergate. Prior to Watergate, Nixon supporters had more faith in the executive branch than supporters of George McGovern, Nixon's opponent in the 1972 election. However, once the scandal unfolded, more Nixon supporters than McGovern supporters lost faith in the executive branch (Zimmer 743). Betrayal played a big role in the loss of support and shaping of public reaction. This effect was especially strong among educated Americans. People with higher levels of education are generally more trusting of others, which made Watergate and the betrayal of Nixon particularly more damaging for them (Zimmer 743-744). Educated voters who supported Nixon likely experienced the greatest drop in trust because they began with higher confidence in the government and then felt more let down when Nixon's actions were exposed. Their political awareness meant they better understood the seriousness of the scandal, intensifying the sense of betrayal. Many political cartoons were released of Nixon, reflecting the public's opinion of him. Fig. 2 reflects the demand for Nixon's resignation, the big hand representing the hand of the public.

Fig. 2. Herblock, Nixon Signing Resignation, August 9, 1974, illustration, https://thenib.com/richard-nixons-impeachment-herblock-cartoons/.

Watergate might not have seemed significant if it were an isolated incident, but it happened right after the Vietnam crisis, which had already begun to create a decline in public confidence regarding the government. Together, Vietnam and Watergate created a broader pattern of government dishonesty, leading to the lowest levels of public trust in government recorded in polling history (Sundquist 453). The state of the country stemmed from the misuse of presidential power. Nixon had already offended the public by prolonging the war despite the public demanding withdrawal, and then he deceived the public with lies and misused power (Sundquist 453). This made the public lose trust in the government altogether.

Nixon's emotional manipulation and lifelong obsession with controlling his public image led to the collapse of the thing he cared most about: his presidency. Throughout his early campaigns, Nixon repeatedly relied on deception, media manipulation, and information control to preserve a carefully constructed public persona. He would do this to win elections. He had multiple controversies along his way to becoming president that he was able to get out of using these tactics. At the same time, the Watergate scandal spiraled out of control because of his deceptive tactics. Had he not lied to the entire nation and manipulated the investigation, he would not have lost the presidency that he worked so hard for. The exposure of his persona as fake caused such a big ripple because it shattered public trust at a time when Americans still largely believed the president was honest and acted in the nation's best interest. The pattern of leaders lying to the nation and manipulating behind the scenes still exists in today's world, and it similarly contributes to public distrust in the government. Public debates over misinformation, media narratives, and questions over the intentions of leaders have remained central in American politics. Nixon's presidency serves as just another example of the lasting consequences when political leaders put self-preservation ahead of the truth.

Works Cited

ABC News. "The Watergate tapes' infamous 18.5-minute gap and Nixon's secretary's unusual explanation for it Rose Mary Woods testified she recorded over some of the tape by mistake." ABC News. https://abcnews.go.com/Politics/watergate-tapes-infamous-185-minute-gap-nixons-secretarys/story?id=47926329.

A&E Television Networks. "President Nixon Refuses to Hand over Tapes." History. Last modified July 16, 2025. https://www.history.com/this-day-in-history/july-23/nixon-refuses-to-hand-over-tapes

Basinger, Scott J., and Brandon Rottinghaus. "Skeletons in White House Closets: A Discussion of Modern Presidential Scandals." Political Science Quarterly 127, no. 2 (2012): 213-39. http://www.jstor.org/stable/41503282.

Bernstein, Carl, and Bob Woodward. All the President's Men. 50th ed. Simon and Schuster Paperbacks, 2022.

Dean, John, H.R. Haldeman, and Richard Nixon. "Transcript of a Meeting Among the President, H.R. Haldeman, and John Dean, on September 15, 1972, at 5:27 to 6:17 PM (First Installment)." Interview. Ford Library Museum. https://www.fordlibrarymuseum.gov/sites/default/files/2024-05/document_transcript_4_2.pdf.

Emery, Fred. Watergate: The Corruption of American Politics and the Fall of Richard Nixon. Touchstone, 1995.

Farrell, John A. Richard Nixon: The Life. Vintage Books, A Division of Penguin Random House, 2018.

Glass, Andrew. "Gap on key Watergate tape revealed: Nov. 21, 1973." Politico. https://www.politico.com/story/2016/11/gap-on-key-watergate-tape-revealed-nov-21-1973-231651.

Herblock. Nixon Signing Resignation. August 9, 1974. Illustration. https://thenib.com/richard-nixons-impeachment-herblock-cartoons/.

Hoover Institution Library & Archives. "Un-Presidented Watergate and Power in America." Hoover Institution Library & Archives. https://www.hoover.org/library-archives/histories/un-presidented.

Jacobs, Lawrence R., and Robert Y. Shapiro. "Presidential Manipulation of Polls and Public Opinion: The Nixon Administration and the Pollsters." Political Science Quarterly 110, no. 4 (1995): 519-38. https://doi.org/10.2307/2151882.

National Archives. "The Watergate Files" [The Watergate Files]. National Archives. Ford Library Museum. https://www.fordlibrarymuseum.gov/exhibits/watergate-files#event-number-1154.

The New York Times. "Transcript of Nixon's Question and Answer Session with A. P. Managing Editors." November 18, 1973.

https://www.nytimes.com/1973/11/18/archives/transcript-of-nixons-question-and-answer-session-with-a-p-managing.html.
The New York Times. "Watergate." April 22, 1973. https://www.nytimes.com/1973/04/22/archives/watergate-aug-29-1972-april-17-1973-watergate-revelations.html.
Nixon, Richard. "'I made clear there was to be no coverup' (April 29,1974)." In Richard Nixon: Speeches, Writings, Documents, by Rick Perlstein. Princeton University Press, 2010.
Nixon, Richard. "'My mother was a saint' (August 9,1974)." In Richard Nixon: Speeches, Writings, Documents, by Richard Milhous Nixon and Rick Perlstein. Princeton University Press, 2010.
Nixon, Richard. "Richard Nixon, Press Conference, Nov 7, 1962." languagelog. https://languagelog.ldc.upenn.edu/myl/RichardNixonConcession.html.
Nixon, Richard. "Senator Nixon's Checkers Speech." watergate.info. Last modified October 12, 2004. https://web.archive.org/web/20041012103913/http://www.watergate.info/nixon/checkers-speech.shtml.
Nixon, Richard Milhous, and Rick Perlstein. Richard Nixon: Speeches, Writings, Documents. Princeton University Press, 2010.
Oelsner, Lesley. "'Pushing of Keys' Caused Tape Gap, Experts Assert." New York Times, June 5, 1974. https://www.nytimes.com/1974/06/05/archives/-pushing-of-keys-caused-tape-gap-experts-assert-panel-rules-out.html.
Pach, Chester. "'Our Worst Enemy Seems to Be the Press': TV News, the Nixon Administration, and U.S. Troop Withdrawal from Vietnam, 1969–1973." Diplomatic History 34, no. 3 (2010): 555-65. http://www.jstor.org/stable/24915901.
Perlstein, Rick. "Watergate Scandal" [Watergate Scandal]. Britannica. Last modified September 16, 2025. https://www.britannica.com/biography/Richard-Nixon.
Reeves, Richard. President Nixon: Alone in the White House. Simon & Schuster, 2002.
Rozell, Mark J. "The Limits of White House Image Control." Political Science Quarterly 108, no. 3 (1993): 453-80. https://doi.org/10.2307/2151699.
Stone, Roger J., and Mike Colapietro. Tricky Dick: The Rise and Fall and Rise of Richard M. Nixon. Skyhorse Publishing, 2017.
Sundquist, James L. "Reflections on Watergate: Lessons for Public Administration." Public Administration Review 34, no. 5 (1974): 453-61. https://doi.org/10.2307/975092.
Thomas, Evan. Being Nixon: A Man Divided. Random House, 2015.
Woodward, Bob, and Carl Bernstein. The Final Days. Simon & Schuster Paperbacks, 2005.
Zimmer, Troy A. "THE IMPACT of WATERGATE on THE PUBLIC'S TRUST in PEOPLE AND CONFIDENCE in THE MASS MEDIA." Social Science Quarterly 59, no. 4 (1979): 743-51. http://www.jstor.org/stable/42860478.

Intermittent Fasting for Adolescent Athletes: A Literature Review By Yahya Dar

Abstract

A newly popular diet, intermittent fasting focuses on constraining eating to certain time periods rather than managing calories or nutrients. It has also increased in popularity among adolescent athletes looking to lose fat and gain muscle to increase athletic performance. This popularity has raised concerns about whether intermittent fasting is a safe and effective diet for teenage athletes. The literature reviewed to address this question consisted of 10 articles, made up of academic papers, government articles, and health and nutrition websites focusing on the physiological effects of the diet. Of these articles, 9 were published from 2008-2025 and one paper was published in 1998, indicating the relevance and increasing popularity of this diet. Few papers conclude that intermittent fasting can be an effective diet, with it being seen as only viable for short term weight loss for healthy teen athletes or for prolonged usage for adults with obesity, but not necessarily adolescents. The existing literature largely shows that intermittent fasting is not a long-term sustainable recommendation for adolescents and can harm their growth and health. They also showed that the main issue with the diet was not the structure but the difficulty to eat sufficient macronutrients and nutrients within the allotted time as, realistically, a teenager might not manage the required meal preparation and planning around their fasting periods appropriately. Overall, intermittent fasting is beneficial to athletic teenagers in certain situations for short periods, but the usage of this diet should be avoided for adolescent athletes to prevent suppression of growth and harm to health.

Introduction

Intermittent fasting is an eating pattern where the timing of the meals is restricted rather than caloric intake or macronutrients being restricted (de Cabo & Mattson, 2019; Zouhal et al., 2020). There are various methods one can employ to implement the fast, including 16/8, the 5:2 diet, Eat-Stop-Eat, and alternate day fasting (Zouhal et al., 2020). The popularity of this diet emerged in 2012, when Dr. Michael Mosley's BBC documentary “Eat, Fast and Live Longer” was broadcast to the public. His documentary showed his journey to losing weight and improving his blood sugar and cholesterol levels all in 5 weeks using the 5:2 method. After this revelation became mainstream, new forms of fasting, such as the 4:3 and 16:8 protocols, were developed, but most of this popularity remained in the United Kingdom. On top of that, high profile individuals, such as CEOs, celebrities, and influencers, adopted and promoted the diet. Social media adopted the diet as a trend reaching millions of posts using #intermittentfasting (Lister et al., 2023; Guan et al., 2025). By 2018, it was the most followed diet in America (IFIC, 2018). It has also gained the attention of the scientific community who published many positive articles that suggest intermittent fasting’s benefits stretch beyond weight loss. Now, the diet has become a popular option for health enthusiasts. The diet is a particularly good diet option for overweight individuals and those with Insulin Resistance or Type 2 Diabetes (de Cabo & Mattson, 2019; Zouhal et al., 2020; Matta et al., 2025).

However, intermittent fasting might not be as promising as it sounds. A study found that most of the weight loss came from a calorie deficit rather than the timing, though restrictions on time did help reduce total calories consumed as suggested by earlier studies (de Cabo & Mattson, 2019). Some concerns with intermittent fasting are that it does not fully optimize results for individuals looking to build or even fully conserve muscle mass during the diet time (Mendonça et al., 2024; Trommelen et al., 2019; Zouhal et al., 2020). Protein synthesis is most optimal when protein is consumed in multiple meals spread throughout the day (Kerksick et al., 2008; Trommelen et al., 2019), consuming at least 0.2-0.5 grams of protein per kilo of bodyweight (Kerksick et al., 2008). Keeping in mind this figure is for maintenance, the protein needed during intermittent fasting would be substantially higher as the body is usually in a state of deprivation during this diet (Trommelen et al., 2019). When used for muscle growth, intermittent fasting becomes challenging as surplus calories, sufficient macronutrients, and a certain degree of meal timing need to be managed.

Despite these challenges, adolescents, particularly athletic adolescents, have gravitated toward intermittent fasting as a way to lose weight and build muscle. However, the fact that female adolescents "require 1,400–2,400 calories per day" and males "1,600–3,200" with some athletic teens who "can require up to 5,000 calories per day" suggest that intermittent fasting may not be optimal for adolescents (Lassi et al., 2017). This review looks at if intermittent fasting is beneficial or detrimental to the health of adolescent athletes.

Methods

The literature review for this study consists of published academic articles and government, health, and nutrition sites that focus on the physiological effects of intermittent fasting with respect to nutrient timing and protein synthesis. Studies selected explored if intermittent fasting has developmental risks or benefits for athletic adolescents. Of the articles selected, 9 were published between 2008 and 2025, and one article was published in 1998. Sources were found through Google Scholar, PubMed, medical journals such as the New England Journal of Medicine, and public access journals like MDPI, Springer, and Biomed Central. The search terms included "intermittent fasting for adolescents," "nutrient timing," "anabolic window," "protein synthesis," "resistance-type exercise," "muscle protein synthesis," "protein metabolism," and "growth." In all, a total of 19 articles were screened, and 9 were chosen. The 9 selected consisted of 4 peer-reviewed research articles, 4 peer-reviewed review articles, and one educational evidence-based clinical guidance resource, though not peer-reviewed. The findings are summarised below.

Results

Benefits of Intermittent Fasting for Adolescents

When looking at diets for adolescents, especially athletes, there must be a balance of sustainability, efficiency, and impact. Intermittent fasting is beneficial for overweight athletes

with an average age of 25 who need to reduce weight to improve performance in their sport (Zouhal et al., 2020). It also helps improve their metabolic health in factors such as insulin sensitivity and cholesterol (de Cabo & Mattson, 2019; Matta et al., 2025; Zouhal et al., 2020). Another benefit of intermittent fasting is improved discipline and corrected habits. If followed correctly, adolescents are less likely to unhealthily snack or eat at late hours, which are habits that are tied to poor health outcomes.

Intermittent fasting also has significant benefits for short term cuts for the general population in weight-based sports such as wrestling and rowing. Other diets, such as the ketogenic diet, similar to intermittent fasting, pose significant benefits such as high fat mass loss with minimal lean mass reduction in adolescents with obesity (Mistretta et al., 2017). If appropriately used, diets like intermittent fasting and ketogenic diets can help the athletes cut their weight in the short term for weigh-ins in sports with weight classes as fat is mostly lost rather than lean mass. These diets can also be used by overweight athletes who need to cut weight for performance rather than weight classes, as this diet particularly helps overweight individuals lose fat mass without affecting lean mass tissue (de Cabo & Mattson, 2019; Willi et al., 1998).

Risks of Intermittent Fasting for Adolescents

On the other hand, intermittent fasting poses significant risks for adolescents, especially those at normal or below normal weights. Individuals who are already healthy could use this diet, but potential risks include impaired metabolic health and insulin production (Matta et al., 2025), along with eating disorders appearing or worsening and insufficient nutrients for proper growth (Kumar et al., 2025). When considering the use of intermittent fasting or the ketogenic diet by average adolescent athletes, prolonged usage of these types of diets can lead to an inhibition of growth. With individuals who are below normal weight, these effects would be far worse and can create irreversible damage. When considering adolescent athletes specifically, the promise of this diet decreases (Zouhal et al., 2020).

Athletes need to optimize their strength, speed, and agility no matter what sport they play. A key factor to building these areas is muscle (Kerksick et al., 2008; Trommelen et al., 2019; Zouhal et al., 2020). To get the most muscle growth, meals need to be planned around training. This includes pre workout meals with carbohydrates for energy during workouts and post workout meals with protein to help with muscle growth during recovery (Kerksick et al., 2008; Trommelen et al., 2019). With most forms of intermittent fasting, getting an adequate amount of food daily is challenging already, and trying to time meals adds another level to it (Zouhal et al., 2020). Along with that, fasting days are fatiguing, and athletes using intermittent fasting during competition days will likely not compete well (Zouhal et al., 2020). Lastly, intermittent fasting, with its often reduced calorie intake, suppresses the mTOR pathway, which in return“impairs insulin production” (Kumar et al., 2025). Insulin, a key hormone that allows energy usage for cells in muscles, when suppressed can lead to reduced muscle growth and hormonal dysfunction, causing teenage bodies to experience reduced performance and even inhibited growth. Overall,

intermittent fasting should not be used for adolescent athletes as it poses significant and long-lasting negative effects to the body if used improperly (Matta et al., 2025; Zouhal et al., 2020).

Similar to intermittent fasting, the ketogenic diet shows the truth behind such restricting diets. When tested in adolescents with obesity, the ketogenic diet helped preserve lean mass while losing predominantly fat (Willi et al., 1998). Realistically, this diet would work only in extreme cases where an adolescent athlete may need short term rapid weight loss for weight-based sports or teen athletes with obesity. Still, this diet is more likely to harm young users as it does not allow for proper nutrition for adolescents to grow. Since this diet imitates the suppressive and low frequency meals in intermittent fasting, it is fair to say that both diets would not be suitable for an average teenage athlete as it is extremely low in calories, preventing crucial growth, inhibiting athletic performance, and harming the individual's body.

Another commonly discussed diet in health spaces is the Mediterranean diet. When compared to the Mediterranean diet, flaws in intermittent food appear clear. In the Mediterranean diet, 3-4 meals are eaten a day consisting mostly of plant-based foods and fish. This ensures adequate nutrients and macronutrients are consumed while lowering the risk of unhealthy eating or overeating, as shown by how “the Mediterranean diet resulted in 30% decreased odds of being overweight or obese” in adolescents (Mistretta et al., 2017). Meanwhile, intermittent fasting usually consists of eating few, protein rich foods that often do not have sufficient nutrients and macronutrients to sustain an active adolescent. Intermittent fasting also keeps the meals within a tight time frame preventing healthy beneficial snacking before physical exercise, unlike the Mediterranean diet where snacking is easy to follow and even encouraged.

Discussion

Intermittent fasting proves to be a situational and experimental diet when compared to meal structures such as traditional, frequent, and nutrient timing meal plans (Kerksick et al., 2008; Zouhal et al., 2020) which can be beneficial for adolescents in some cases but definitely not recommended as the best diet for adolescent athletes in general. One situation where intermittent fasting may be considered the best option for a meal plan for both adolescents and the general population is when significant fat loss is required in order to prevent health conditions like diabetes (de Cabo & Mattson, 2019; Matta et al., 2025; Zouhal et al., 2020).

One benefit that can be derived from following intermittent fasting for adolescents in general is that it teaches them to be responsible with their food choices and shows them how to follow a set plan. These factors show what benefits intermittent fasting can have for athletes who need to cut weight, improve metabolic health, and build discipline. However, intermittent fasting falls short as a diet for athletes in other ways, as it is hard to manage, reduces nutrient optimization, and makes it difficult to have healthy quantities of food for most populations (Trommelen et al., 2019; Zouhal et al., 2020). Adolescent athletes are growing and need energy to perform at their best (Matta et al., 2025). Moderate weight loss is better planned with frequent meals to reduce cravings and prevent the teen from overeating. When bulking to add weight or

muscle, the nutrient timing approach where meals are structured around workouts rather than pre-set fasting periods helps optimize muscle protein synthesis (Kerksick et al., 2008; Trommelen et al., 2019). When maintaining weight, traditional meals provide a simple and sustainable structure that is already implemented by many adolescent athletes.

Adolescents considering this diet should be informed about the risks of intermittent fasting. Recent studies have suggested that rather than improving insulin sensitivity, due to the reduced calorie intake, the mTOR pathway is suppressed, and "the reduction in mTOR activity impairs insulin production" causing lower insulin sensitivity (Kumar et al., 2025, p. 2). This increases the chance of diabetes and other issues related to glucose levels. Also, the timing of intermittent fasting is harmful particularly to athletic adolescents as it often interferes with school timing. By fasting in school, energy levels are lower as the day progresses, hunger increases, and by the time practice begins, usually after school, the individual is exhausted and not able to perform at an optimal level. Intermittent fasting may be an effective weight loss diet, but its practicality fails to include an adolescent's growth, especially ones in athletics (Kumar et al., 2025). Overall, intermittent fasting may be a viable option for overweight individuals to lose fat (de Cabo & Mattson, 2019; Matta et al., 2025; Zouhal et al., 2020), but other diets with more frequent meals pose greater benefits for less negative effects.

When looking at what may be more optimal for adolescents, balance is key. Athletes need to have around "60% carbohydrates, 20% protein and 20% fat" in their diet and have sufficient vitamins and minerals (Nairn, 2025). This can be easily achieved, most arguably through either the traditional three meals a day or a modified version with 1-3 additional meals with the needed compensations. First, the traditional three meals a day allows the adolescent to eat the needed nutrients easily throughout the day without having to meticulously check every meal's macronutrients. It also allows for more leeway as missing a target of a specific number of grams of protein in one meal can be compensated for in a later meal. When adjusting one's diet to 4-6 meals a day, the individual has much more flexibility but must be more aware of what they are eating. For most healthy diets, meals should be intended to relieve hunger without causing fullness, allowing for more satiating meals that prevent overeating (Nairn, 2024). A key issue in meal plans focused on fasting for hours is the risk of binge eating after. Diets with more frequent and adjustable meals also allow an individual to choose the meals and what they have significantly more freely as they have the choice of how many meals they eat, how often they eat, and how much they eat. This makes the diet enjoyable and interactive, contrasting other repetitive and hard to follow meal structures. Overall, having more frequent meal timings allows for easier application, prevention of binging, and more leeway for mistakes.

Works Cited

de Cabo, R., & Mattson, M. P. (2019). Effects of Intermittent Fasting on Health, Aging, and Disease. *New England Journal of Medicine*, *381*(26), 2541–2551. https://doi.org/10.1056/nejmra1905136

Guan, Q., Mai, W., Qiu, Z., & Zuo, Y. (2025). Is time-restricted eating a healthy choice to lose weight? Investigating by qualitative analysis of Instagram posts and systematic reviews with meta-analysis. Digital Health, 11. https://doi.org/10.1177/20552076251360911

IFIC. (2018). "One-Third of Americans Are Dieting, Including One in 10 Who Fast … While Consumers Also Hunger for Organic, "Natural" and Sustainable - IFIC." IFIC, 13 June 2025, ific.org/media/one-third-of-americans-are-dieting-including-one-in-10-who-fast-while-consumers-also-hunger-for-organic-natural-and-sustainable/.

Kerksick, C., Harvey, T., Stout, J., Campbell, B., Wilborn, C., Kreider, R., Kalman, D., Ziegenfuss, T., Lopez, H., Landis, J., Ivy, J., & Antonio, J. (2008). International society of Sports nutrition position stand: Nutrient timing. In Journal of the International Society of Sports Nutrition. https://pubmed.ncbi.nlm.nih.gov/18834505/

Kumar, D., Gupta, V., Rajni Tanwar, & Bansal, V. (2025). Unveiling the Hidden Dangers of Intermittent Fasting in Adolescents: A Call for Awareness. PubMed, 40(11). https://doi.org/10.1111/jgh.70077

Lassi, Z., Moin, A., & Bhutta, Z. (2017). Nutrition in Middle Childhood and Adolescence. PubMed; The International Bank for Reconstruction and Development / The World Bank. https://www.ncbi.nlm.nih.gov/books/NBK525242/

Lister, N. B., Melville, H., & Hiba Jebeile. (2023). What adolescents see on <scp>instagram</scp>: <Scp>content</scp> analysis of #intermittent fasting, #keto, and #lowcarb. Nutrition & Dietetics, 81(3). https://doi.org/10.1111/1747-0080.12853

Lopez, H., Landis, J., Ivy, J. L., & Antonio, J. (2008). International Society of Sports Nutrition position stand: Nutrient timing. Journal of the International Society of Sports Nutrition, 5(1). https://doi.org/10.1186/1550-2783-5-17

Matta, L., Weber, P., Erener, S., Walth-Hummel, A., Hass, D., Bühler, L. K., Klepac, K., Szendroedi, J., Guerra, J., Rohm, M., Sterr, M., Lickert, H., Bartelt, A., & Herzig, S. (2025). Chronic intermittent fasting impairs β cell maturation and function in adolescent mice. *Cell Reports*, *44*(2), 115225. https://doi.org/10.1016/j.celrep.2024.115225

Mendonça, M. L. M., Carvalho, M. R., Romanenghi, R. B., Santos, D. S. D., Filiú, W. F. O., Pagan, L. U., Okoshi, K., Okoshi, M. P., Oliveira, R. J., Oliveira‐Junior, S. A., & Martinez, P. F. (2024). Impact of combined intermittent fasting and high‐intensity interval training on apoptosis and atrophy signaling in rat fast‐ and slow‐twitch muscles. *Physiological Reports*, *12*(15). https://doi.org/10.14814/phy2.16181

Mistretta, A., Marventano, S., Antoci, M., Cagnetti, A., Giogianni, G., Nolfo, F., Rametta, S., Pecora, G., & Marranzano, M. (2017). Mediterranean diet adherence and body

composition among Southern Italian adolescents. Obesity Research & Clinical Practice, 11(2), 215–226. https://doi.org/10.1016/j.orcp.2016.05.007

Nairn, R. (2024, October 3). Nutrition for Athletes: What to Eat Before a Competition. Hopkinsmedicine.org. https://www.hopkinsmedicine.org/health/wellness-and-prevention/nutrition-and-fitness/nutrition-for-athletes-what-to-eat-before-a-competition

Trommelen, J., Betz, M. W., & van Loon, L. J. C. (2019). The Muscle Protein Synthetic Response to Meal Ingestion Following Resistance-Type Exercise. Sports Medicine, 49(2), 185–197. https://doi.org/10.1007/s40279-019-01053-5

Willi, S. M., Oexmann, M. J., Wright, N. M., Collop, N. A., & Key, L. L. (1998). The Effects of a High-protein, Low-fat, Ketogenic Diet on Adolescents With Morbid Obesity: Body Composition, Blood Chemistries, and Sleep Abnormalities. Pediatrics, 101(1), 61–67. https://doi.org/10.1542/peds.101.1.61

Zouhal, H., Saeidi, A., Salhi, A., Li, H., Essop, M. F., Laher, I., Rhibi, F., Amani-Shalamzari, S., & Ben Abderrahman, A. (2020). Exercise Training and Fasting: Current Insights. Open Access Journal of Sports Medicine, Volume 11(2), 1–28. https://doi.org/10.2147/oajsm.s224919

How the Stages of Adolescence Affect Personality Traits By Mariela Vargas

Abstract

The adolescent stage is a particularly challenging period in development. This stage is challenging because it is marked by changes to both the body and the mind. This article reviews four articles which discuss the personality types in both childhood and adolescence, known as the little six and big five, and how these personality types change through adolescence. All four articles report the personality types (extraversion, agreeableness, conscientiousness, neuroticism, openness to experience, and activity), how they change, and how gender relates to personality traits.

Introduction

The adolescent stage is a period in development known for being difficult. This stage is particularly challenging because it is marked by changes to both the body and the mind. Adolescents also gain additional knowledge and extraversion during this time. Teens are acutely aware of their doings and if these behaviors are socially acceptable. This hyperawareness of behavior for oneself and others can lead to emotional struggles and altering how adolescence presents itself. In the words of John Milton, "the mind is its own place, and in itself can make a heaven of hell, a hell of heaven." Essentially, adolescents often get trapped in a bubble, designed by their own mind, and experiences, often morphing the bubble or personalities.

One of the only things that adolescents have control of is their own mind. This becomes their biggest asset, as the mind controls actions and personalities that eventually influence development of friend groups through finding individuals with similar personalities. Psychologists categorize children personalities using the little six, the categorical hierarchy of children's personalities are extraversion which is a personality characterized by sociability, assertiveness, and energy level; agreeableness which is a personality characterized by compassion, politeness, and trust in others; conscientiousness which is a personality characterized by organization, industriousness, and reliability; neuroticism which is a personality characterized by anxiety, depressiveness, and emotional volatility, and openness to experience which is a personality characterized by intellectual curiosity, creativity, and aesthetic sensitivity; and activity which is a personality characterized by physical energy and motor activity (Soto and Tackett).

As the children transition into an adolescent so can their personality which can bring in a new categorical hierarchy of personalities into the big five. The big five is an adult hierarchy of the core personalities. The first personality type in the big five is extraversion which is characterized by sociability, assertiveness, and energy level. The second is agreeableness which is characterized by compassion, politeness, and trust in others. The third is conscientiousness which is characterized by organization, industriousness, and reliability. The fourth is neuroticism which is characterized by anxiety, depressiveness, and emotional volatility. Finally, openness to experience is characterized by intellectual curiosity, creativity, and aesthetic sensitivity (Soto and

Tackett).

Using the categorical hierarchy psychologists have conducted studies to examine gender differences in personality traits. These studies have shown that across adolescence, girls exhibit more consistent changes in personality including — increases in self-reported openness and conscientiousness, and mixed patterns for extraversion. By contrast, boys show fewer changes, primarily declines in extraversion and openness. In adulthood, fathers' extraversion, agreeableableness, and emotional stability declined, while mothers' self-reported emotional stability and conscientiousness increased. Causes of these changes include personal experiences which may increase or decrease personality characteristics (Napolitano et al.).

Therefore, the aim of this review is to answer the question, how can the stages and events of adolescence affect personality traits? The goal of this work is to better understand how the transition from childhood to adolescence can alter or effect changes in personality. In other words, how the personalities transition from the little six to the big five.

Methods

A literature review was conducted by identifying peer reviewed studies that examine personality and mental health through adolescence. Literature searches were conducted on Google Scholar between March and May 2026. Search terms include 'personality types,' 'personality traits among teenagers,' 'personality traits in childhood and adolescence,'and 'transactions between life events and personality traits across the adult lifespan.' Research articles were included if they met the following criteria: 1) examined personality traits in adolescence, 2) were published between 2015 and 2026, and 3) written in the English language. Articles were excluded if 1) content did not focus on personality traits during adolescence (e.g., in childhood or adulthood only), and/or 2.) were published before the year 2015.

Discussion

Personality Traits in Childhood and Adolescence: Structure, Development, and Outcomes

A deep analysis conducted on adult personality development supports two key principles. One is the cumulative-continuity principle which states that personality becomes increasingly stable across adulthood. The second conclusion is the maturity principle which states that most people become more agreeable, conscientious, and emotionally stable (i.e., less neurotic) with age (Roberts and DelVecchio). Therefore the cumulative-continuity principle highlights how personality stabilizes in adulthood and the maturity principle emphasizes how personalities generally become less neurotic overall. Other studies have examined if these principles are applied to childhood and adolescent personality development (Soto et al.). According to prior work, the average rank-order stability of personality traits steadily increases from infancy through childhood, adolescence, and adulthood. The same work shows that the maturity principle does not fit into adolescent personality development. Instead, findings point to the disruption hypothesis, which proposes that the biological, social, and psychological

transitions from childhood to adolescence are accompanied by temporary dips in some aspects of personality maturity. Researchers found that the mean levels of agreeableness, conscientiousness, and openness decline from late childhood to early adolescence, then a rapid increase from late adolescence to early adulthood, and finally a slow incline through adulthood and middle ages (Van den Akker et al.). Therefore, we can conclude that only one principal is seen through the average life span, and which personality types are presented in the average life span.

Big Five Personality Development in Adolescence and Adulthood

In a gender specific report on mid life personality development, both males and females were found to longitudinally decrease in extraversion and increase in emotional stability. By contrast, only males were found to decrease in conscientiousness and openness, while females only increased in agreeableness (CostaJr. et al.). A cross-sectional study using internet reports, showed a sharp increase in agreeableness among adult women resulting in smaller gender differences in emotional stability with age. These results also showed that extraversion slightly decreased in women and a small increase for men (Srivastava et al.). In sum, these two works support that personality traits change over time and these changes may be gender-specific.

Limitations & Future Directions

The findings from this literature review should be considered in the context of its limitations. First, we sought to understand the personality types before and after adolescence, however literature assessments on categorizing the big five and little six were limited. This may be due to limited prior longitudinal investigations or limited published findings due to lack of identifiable patterns. Therefore, future research should focus on the categorisation of personality traits through adolescence and adulthood to elucidate clear patterns.

Conclusions

In conclusion, the goal of this literature review was to understand how the little six childhood personality traits convert into the adult big five before personality traits through the adolescent period. Specifically, four research articles were examined to further review the transformation of personality across the life span. The results of the included research suggest that the little six personality traits transform into the big five with predicted outcomes More research is needed to identify patterns and predicated outcomes.

Works Cited

CostaJr., Paul T., et al. “Personality at Midlife: Stability, Intrinsic Maturation, and Response to Life Events.” Assessment, vol. 7, no. 4, Dec. 2000, pp. 365–78. https://doi.org/10.1177/107319110000700405.

Napolitano, Christopher M., et al. “Social, Emotional, and Behavioral Skills: An Integrative Model of the Skills Associated With Success During Adolescence and Across the Life Span.” Frontiers in Education, vol. 6, June 2021. https://doi.org/10.3389/feduc.2021.679561.

Roberts, Brent W., and Wendy F. DelVecchio. “The Rank-Order Consistency of Personality Traits from Childhood to Old Age: A Quantitative Review of Longitudinal Studies.” Psychological Bulletin [US], vol. 126, no. 1, 2000, pp. 3–25. https://doi.org/10.1037/0033-2909.126.1.3.

Soto, Christopher J., et al. “Age Differences in Personality Traits from 10 to 65: Big Five Domains and Facets in a Large Cross-Sectional Sample.” Journal of Personality and Social Psychology [US], vol. 100, no. 2, 2011, pp. 330–48. https://doi.org/10.1037/a0021717.

Soto, Christopher J., and Jennifer L. Tackett. “Personality Traits in Childhood and Adolescence: Structure, Development, and Outcomes.” Current Directions in Psychological Science, vol. 24, no. 5, Oct. 2015, pp. 358–62. https://doi.org/10.1177/0963721415589345.

Srivastava, Sanjay, et al. “Development of Personality in Early and Middle Adulthood: Set like Plaster or Persistent Change?” Journal of Personality and Social Psychology [US], vol. 84, no. 5, 2003, pp. 1041–53. https://doi.org/10.1037/0022-3514.84.5.1041.

Van den Akker, Alithe L., et al. “Mean-Level Personality Development across Childhood and Adolescence: A Temporary Defiance of the Maturity Principle and Bidirectional Associations with Parenting.” Journal of Personality and Social Psychology [US], vol. 107, no. 4, 2014, pp. 736–50. https://doi.org/10.1037/a0037248.

Neurological Symptom Burden and Repetitive Head Impact Exposure in Figure Skaters: A Cross-Sectional Survey By Amal Israilova

Abstract

Background

Figure skating contains rotational forces, high-impact falls, and deceleration and acceleration movements, which may make up either concussive or subconcussive neurological trauma. Neurologically speaking, the implication of repetitive head contact with the ice in figure skating is still largely overlooked, even though the awareness of traumatic brain injury (TBI), particularly in contact sports, is expanding. It has been speculated that the consistent exposure to subconcussive forces initiates the pathways for inflammasome activation and the hyperphosphorylation of tau protein, pointing to a serious issue concerning the danger of neurodegenerative problems.

Objectives

This study aimed to (1) describe the neurological symptoms that competitive and recreational figure skaters have self-reported, (2) analyze the history of concussions and the repetitive head-ice contact frequency, (3) evaluate help-seeking actions after head impact, and match the parental observations with the skaters' self-reports through matched pairs.

Methods

A cross-sectional survey was distributed to 99 figure skaters (ages 6-54; 93.9% female) and 73 parents or guardians throughout national and international figure skating communities. The skaters completed a neurological symptom inventory consisting of eight parts (0-4 Likert scale). Parents completed a parallel observer-based inventory of five items. Thirty-two matched skater-parent pairs were found through joint unique codes between each pair for cross-information data analysis.

Results

35.4% of skaters reported a diagnosed concussion history, with 26.3% reporting falls involving head contact with the ice on a monthly basis. The mean total symptom burden was 10.43/32, with headaches (mean of 1.72/4) and changes in mood (mean of 1.46/4) being most prevalent. Disorientation following movements of rotation was reported at an average of 1.45/4. 45.5% of all figure skaters reported dismissing head pain after impact, which increased to 62.9% amid the skaters with concussion histories. Within the skater-parent pairs, 64.3% of parents of concussed skaters reported noticing changes in behavior, as opposed to 16.7% of parents of non-concussed skaters. Dizziness observed and reported by parents (0.67 vs 0.36) and forgetfulness (0.62 vs 0.24) were higher in children that had experienced head contact. The percentage of parents who seeked out medical help after noticing changes was 37%

Conclusions

Figure skaters undergo a significant impact or burden of neurological symptoms, including those without any formal concussion histories, indicating that the subconcussive contact may contribute greatly to symptom reflection. The elevated rates of dismissing pain and postponing medical assistance demonstrate substantial obstacles when it comes to neuroconservation and protection in this demographic. The findings demonstrate the necessity of instilling concussion sport-specific education and neurological surveillance procedures in figure skating.

1. Introduction

Over the past twenty years sport-related traumatic brain injury (TBI) has become one of the most urgent health complications. Although the majority of current and existing works of literature have been concentrating on high contact sports like soccer, American football, boxing, and ice hockey, accumulating data indicates that repetitive subconcussive contact, including events of biomechanical head impact and rapid motion or acceleration that are below the clinical benchmark for diagnosis of concussions, may individually play a role in gradually causing neurologic impairment (McKee et al.; Bailes et al.)

As a sport that is practiced by approximately 700,000 registered skaters and 50 million athletes across the world, figure skating is composed of specialized and atypical biomechanical features and movement patterns, which differentiates itself from the most well-known or commonly researched collision and impact sports. Competitive figure skaters constantly apply centrifugal forces to the head and cervical spine by executing rotational spin and twizzle movement patterns or sequences, perform jumps consisting of multiple revolutions in the air along with sustaining landing forces of sometimes over six times their own bodyweight, and absorb the impact from falls directly involving head contact on the ice at around 5-15 kilometer per hour (Hack et al.; Dubravcic-Simunjak et al.). Other variables of possibilities of injury, specifically through throws, lifts, and specific body placement or contact with each partner, are also introduced in both the ice dance and pairs disciplines.

Even with the level of injury risk or susceptibility that skaters are routinely exposed to, the sport of figure skating has been acquiring much less focus in the sports neurology and literature field. The very few articles that have been published generally concentrate on neurological data/results mostly not present in or from peer-reviewed documents, and they usually simply look at musculoskeletal wounds. This rift raises a cause for concern, especially given the growing emergence of mechanism-focused, fundamental research affecting and triggering the activation of inflammasomes, particularly the NLRP3 pathway. The inflammasome activation creates serious repercussions from top to bottom of mild yet consistently present traumatic brain injury (TBI), which could possibly result in tau protein hyperphosphorylation and therefore the beginning of and main molecular indicators of chronic traumatic encephalopathy (CTE) and Alzheimer's disease (Stancu et al.; Heneka et al.).

In addition, figure skating brings in factors particular to neurocognitive and developmental susceptibility, since the sport is mainly made up of adolescent and female athletes. There are specific patterns of extended period of symptoms, recuperation from concussions, and increased sensitivity to recurring impact and injury as opposed to adults that take place within the brains of pediatric and adolescent minds (Novak et al.; McCrory et al.). A call for comprehensive neurological evaluation or analysis is critical given the junction of young female athlete populations who may still be in the process of physical and mental development, the frequent risk of impact and injury, and a sport culture that may desensitize the feeling of pain and disinclined medical statements or the sharing of health information.

The study was intended to confront this limitation through the conduction of a dual-survey cross-sectional design and gather data regarding neurological symptoms from figure skaters and from their guardians/parents in tandem. The implementation of matched pairs gave way for multi source verification, which is a strength, procedurally speaking, that was not mentioned earlier in this populace. The study hypothesized that (1) a large portion of the skaters would claim to have had neurological symptoms at some point that match with the frequent injury/impact vulnerability; (2) increased symptom burden would be linked to history of concussions; (3) behaviors of dismissing or not addressing pain would be present, specifically in athletes who already had received concussion diagnoses in the past; and (4) parents' observations would verify the skaters' self-reports and disclose overlooked or understated aspects of the skaters' behaviors and symptoms through the inconsistency in responses between the skater and their matching parent/guardian pair reports.

2. Methods

2.1 Study Design and Participant Recruitment

The study implemented a cross-sectional observational survey that was administered online through Google Forms between December 2025 and January 2026. Participants were selected through national and international figure skating club communication networks and channels, coaching groups and associations, and social media platforms focused on competitive and professional figure skating communities.

The criteria of eligibility for the skater survey included: current or former involvement in figure skating at any competition level; 6 years of age or older; and capable of reading and/or completing the survey independently (with parents offering some assistance for skaters under 10 years of age). The parent/guardian survey was reserved for the parents or guardians of the current or former figure skaters who were capable of documenting their child's skating history and behaviors/changes in behavior.

An essential feature utilized in the process was the common unique 4-6 character alphanumeric codes that the participants (figure skaters) and their parents/guardians chose together before completing their respective surveys. The codes allowed for the matching of pairs

without having to collect anyone's private or personal information to identify them, therefore protecting the participants' privacy and facilitating cross-survey comparison.

2.2 Survey Instruments

The skater survey was composed of five sections: (A1) Demographic and training history (age, gender, years skating, weekly training hours, competitive level, and preventative/protective exercise behaviors); (A2) Concussion history, number of diagnosed concussions, management, and time off skating; (A3) Monthly frequency of falls resulting in head-ice contact; (A4) An 8-item neurological symptom inventory based on a four-point Likert scale (0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, 4 = Always) for the past 6 months; and (A5) Pain dismissal behaviors, help-seeking timelines, and frequency of migraines. The symptom inventory evaluated headaches, dizziness or loss of balance, memory issues/difficulties, sensitivity to light or noise, slower reaction time/fatigue, difficulty with concentration, changes in mood/irritability, and disorientation after jumps or spins.

The parent/guardian was composed of five sections: (B1) Demographic information about the skater child; (B2) History of observed head contact on the ice and changes in behavior; (B3) A five-item observer-rated symptom inventory (same 0-4 Likert scale) for parent-observed dizziness, complaints regarding headache(s), forgetfulness, mood swings/changes, and fatigue/confusion; and (B4) Time elapsed before medical assistance was sought out and protection or recovery strategies against impact/injury initiated.

2.3 Data Analysis

Quantitative information was collected through descriptive statistics, such as standard deviations, means, and frequency distributions. The primary between-group comparison analyzed neurological symptom burden (total score, range 0-32) between skaters with and without a diagnosed concussion history. For each of the eight individual symptom domains, the concussion positive and concussion negative groups had their average/mean Likert scores compared.

In the parent/guardian survey, parents of skaters with reported head-ice impact as opposed to those without had their symptom mean scores compared. Matched couplet pair analysis was used to put the parental documentation of changes in behavior (binary yes/no) with skater-documented concussion history side by side for comparison, investigating the cross-information agreement in responses among the 32 identified matched pairs. The statistical significance thresholds were set at $p < 0.05$ through the chi-square tests for categorical variables and independent tests or samples comparisons for continuous variables due to the descriptive and investigative constitution of the study. All inductive or inference statistics are documented as indicative of certain trends. However, the findings should be treated with caution, as they are preliminary and there is a limited sample size.

Every piece of data was anonymized before completing the analysis. One participant (ID: Bs67) was not included in the present analyses because of free-text spaces/entries without

the completion of responses needed, indicating the respondent's failure to fulfill the requirements of the survey. Every other response was completed entirely and thus was later analyzed.

2.4 Ethical Considerations

Participation was completely anonymous and voluntary. None of the participants' personally identifiable information was collected. The only linking mechanism to match the individual skater with their parent/guardian in matched pairs across the surveys was the shared code system. Informed consent was acquired through the introduction message for each survey, which explained the purpose of the study, use of data, and the voluntary composition of the study and participation. The survey introduction for underage/minor participants required parental consent for their child to participate.

3. Results

3.1 Participant Characteristics

There were 99 figure skaters who completed the survey. The ages ranged from 6 to 54 years of age (mean of 18.7 years, standard deviation approximated at about 7.2). The sample was mostly made up of female athletes (number (n) = 93 female, 93.9%; n = 6 male, 6.1%). The average/mean years of skating experience was 11.2 years (ranging 1.5-40). The following represents the competitive level distribution: Senior/Elite (including collegiate, international, and national level skating) was 29.3%; Junior level was 25.3%; Intermediate/Novice/Pre-Silver was 33.3%; and Recreational and other levels was 12.1%. The parent survey resulted in 73 responses, with the documented skaters' ages ranging from 6 to 51 years (average/mean around 15.8 years) and skating experience of 3-18 years.

Participant characteristics are summarized in Table 1

Variable	**Skater Survey (n = 99)**	**Parent Survey (n = 73)**
Age, average/mean (range)	18.7 years (6-54)	Child: 15.8 years (6-51)
Gender (Female percentage)	93.9%	N/A (child-reported)
Years of skating, average/mean (range)	11.2 years (2-40)	11.1 years (3-18)
Competitive level	Senior/Elite: 29.3%; Junior: 25.3%; Intermediate/Novice: 33.3%; Recreational: 12.1%	N/A
Concussion history (skater)	35.4% (n = 35)	Head contact documented: 65.8% (n = 48)
Gender breakdown	Female: 93; Male: 6	N/A

Table 1: Participant Demographics: Skater and Parent Survey Samples

3.2 Concussion History and Head-Ice Contact

35 skaters (35.4%) out of the 99 skaters who participated reported having a history of at least one diagnosed concussion. Within the concussion group, the amount of prior concussions ranged from 1 to 5. The most common approach to concussion management was the skater's removal or time off from skating recommended by a physician (the response most often/frequently reported), with parents/guardians and coaches also being the forces/factors of the skater's removal. The time of recovery ranged from less than a week to over three weeks. In particular, one of the participants gave a description of a historical concussion that took place before any official concussion regulation or guideline was established, which reveals how concussion management and awareness has shifted or evolved over time in the sport

26 out of the 99 skaters (26.3%) documented at least one fall every month that led to head contact with the ice. Within the parent survey group, 48 out of 73 parents (65.8%) reported that their child had at least one experience of falling and making ice contact with their head, which is a significantly greater statistic than that of the skaters' self-report, which may indicate potential underreporting in the skater survey.

3.3 Neurological Symptom Burden

The average/mean neurological symptom burden among 99 skaters was 10.43 out of a potential 32 (demonstrating an average symptom response of about 1.3/4 over all eight categories). Headaches (mean of 1.72/4), changes in mood or irritability (mean of 1.46/4), disorientation after jumps or spins (mean of 1.45/4), and difficulty with concentration (1.42/4) were the most prevalent symptoms, while memory difficulties (mean of 1.11/4), sensitivity to light/noise (mean of 0.87/4), and fatigue (mean of 1.08/4) were not documented as much but still at significant rates.

Unlike the main hypothesis, skaters who had a history of diagnosed concussion did not present significantly higher total symptom burden (mean of 9.83) as opposed to the skaters without a history of diagnosed concussion (mean of 10.77). This result indicates that the reflection of neurological symptom in the sport of figure skating may not be restricted or limited to just the athletes with official/documented concussive experiences and events, and that frequent mechanical, subconcussive strain or load may work as an individual contributor to symptom burden in skaters who do not have official diagnoses of concussion. The pattern demonstrated matches with newer, developing literature focusing on the subconcussive risk and vulnerability in other sports requiring rotational tension and loading (Montenigro et al.).

There were 20 skaters (20.2%) who converged at the baseline of high symptom burden (total score $\geq$ 16/32, suggesting average Likert response of 2+ among categories). Seven had a history of diagnosed concussion while 13 of the 20 high-burden responders did not, which reinforces the idea that the symptom burden is not influenced by formal diagnosis of concussion(s) in the sample.

Mean neurological symptom scores by concussion history are exhibited in Table 2

Neurological Symptom	**Concussion History (n = 35, mean)**	**No Concussion History (n = 64, mean)**	**All Skaters (n = 99)**
Headaches	1.71	1.73	1.72
Dizziness/Loss of Balance	1.17	1.38	1.30
Memory Difficulties	1.17	1.08	1.11
Sensitivity to Light/Noise	0.80	0.91	0.87
Slower Reaction Time/Fatigue	0.89	1.19	1.08
Concentration Difficulty	1.37	1.45	1.42
Changes in Mood/Irritability	1.46	1.47	1.46
Disorientation After Jumps/Spins	1.26	1.56	1.45
Total Symptom Burden (maximum 32)	9.83	10.77	10.43

Table 2: Neurologic Symptom Inventory: Skater Survey (0-4 Likert Scale, Past 6 Months)

3.4 Pain Dismissal and Help-Seeking Behaviors

Dismissing pain after head contact with the ice was very common. 45.5% (n = 45) out of the total 99 skaters reported dismissing head pain after head-ice contact. This rate was significantly increased in skaters who had a history of concussion(s), 62.9% (n = 22 out of 35) of skaters who ignored/dismissed pain. There were several participants who recounted waiting for the pain or dizziness to resolve on its own, coming back to practice and skate more the same day or the next day, and downplaying symptoms to coaches and parents to prevent being removed from the ice.

Within the parent/guardian group, 63.0% (n = 46 out of 73) documented that they never went to seek out medical help after noticing behavioral changes in their child. The percentage of parents who looked for medical examination within the same day was 17.8% (only 13 parents), while only 12.3% (9 parents) within 1-6 hours, and 6.8% waited for more than one day (5 parents). These results suggest prevalent inadequate or insufficient usage of medical supplies and assistance after neurologically significant or alarming incidents within this population. It also indicates a general/widespread disparity in reaction/response guidelines following concussions at both the recreational and competitive training and club level.

Parent-observed symptom scores by head-contact history are shown in Table 3

Parent-Observed Symptom	**Head Contact Group (n = 48, mean)**	**No Head Contact Group (n = 25, mean)**
Dizziness/Imbalance	0.67	0.36
Complaints about Headaches	1.02	0.84
Forgetfulness	0.62	0.24
Mood Swings/Irritability	1.38	1.12
Increased Confusion/Fatigue	0.83	0.44

Table 3: Parent-Observed Symptoms: Head Contact vs. No Head Contact On the Ice

Pain-dismissal and help-seeking behaviors are summarized in Table 4

Time to Seek Medical Assistance	**Skaters (Percentage That Dismissed Pain)**	**Parents (Percentage That Sought Out Care)**
Dismissed/Never sought	45.5% dismissed pain in general 62.9% among concussed skaters	63.0% never sought medical help
Within 24 hours or same day	Minority sought help	17.8% (Within <1 hour or same day)
1-6 hours after incident	N/A	12.3%
More than 1 day later	N/A	6.8%

Table 4: Pain Dismissal and Medical Help-Seeking Behaviors

3.5 Parent-Observed Neurological Symptoms

The parents of skaters who documented head contact with the ice exhibited steadily higher observer-evaluated symptom scores over all five categories as opposed to parents of skaters who did not document head-ice contact. The greatest absolute differences were noticed in forgetfulness (0.62 vs. 0.24; $\Delta = 0.38$), confusion/fatigue (0.83 vs. 0.44; $\Delta = 0.39$), and dizziness (0.67 vs. 0.36; $\Delta = 0.31$). The most commonly, high-frequency parent-observed symptom in both groups was mood swings (1.38 and 1.12, respectively)

Parent-reported qualitative observations expanded on the quantitative results. Parents explained the decline in their child's processing speed and cognitive function, in general, after concussive incidents ("doesn't assimilate information as quickly"), lingering emotional instability "the following year, my child was not herself"), decline in academic performance and concentration ("increased issues paying attention; has had issues retaining information from classes"), and fluctuations in motivation ("she has been working slower and has been less motivated for everything")

3.6 Matched-Pair Analysis

Through shared unique codes, 32 matched skater-parent pairs were recognized. From these 32 matched pairs, 14 pairs had skaters with concussion history and 18 pairs had skaters without concussion history. Within the matched pairs with concussion history, 64.3% of parents documented noticing changes in their child's behavior. On the other hand, only 16.7% of parents within the non-concussed matched skater-parent group reported observing changes in behavior, in which the difference quadrupled, allowing for cross-information verification regarding the connection between concussion history and noticeable effects in behavior.

Notably, unison or consistency was found in 21 cases (65.6%) through the direct comparison of skater-reported history of concussion(s) and parent-reported head impact within the 32 matched skater-parent pairs, while inconsistency was found in 11 cases (34.4%). Parents documented head contact that the skater did not consider or categorize as a concussion in almost all the cases with inconsistency between the matched skater and parent reports. This specific pattern is in line with or matches the common inclination for athletes to both minimize and miscalculate the neurologic concern/effect of head impacts and downplay or underreport incidents of pain and or concussions, which is a general occurrence widely researched/reported in elite sport groups (Kerr et al.)

Results from the matched skater-parent pairs are summarized in Table 5

Matched Pair Variable	**Concussed Skaters (n = 14 matched pairs)**	**Non-Concussed Skaters (n = 18 matched pairs)**
Mean/Average skater symptom burden	9.71/32	10.22/32
Parents who noticed changes in behavior	64.3%	16.7%
Skater-parent unison/agreement regarding head contact on ice	Approximately 66% agreement/unison within all 32 pairs	N/A

Table 5: Matched-Pair Analysis of 32 Skater-Parent Dyads

4. Discussion

4.1 Principal Findings

The study gives the first comprehensive analysis in an interdisciplinary figure skating athlete sample of neurological symptoms that are self-reported and behaviors of seeking medical assistance through a verified multi-informant outline. There are multiple principal findings from the data that have significant consequences for athlete's health, sports medicine regulations, and coaching methods.

To start, neurological symptoms in this population are both present and quantifiable at a likely elevated level or degree than that of what would be predicted in the general group of people of similar age and gender. About one-third of the highest possible score, the mean total symptom burden of 10.43/32, among athletes with various histories of concussions indicates that

the repetitive spin-driven rotational, centrifugal forces/resistance and subconcussive falls, including the figure skating atmosphere itself, may play a role in creating a constant neurological symptom burden independent of diagnosed concussions

In addition, another significant finding is that there is no plain linear relationship between history of diagnosed concussion(s) and symptom burden. Skaters without any concussions documented slightly elevated total symptom burden (10.77) than skaters with concussions (9.83), and most of the participants with the most significant impact (13 out of 20 respondents) reported not having a diagnosed concussion. This contradictory result may exhibit multiple overlapping mechanisms occurring at the same time. One possible mechanism may be that exposure to repetitive subconcussive contact accumulates over time in athletes without any concussions, which results in neurological symptoms that can be measured even without the presence of an official diagnosis. Another possibility is a selection bias in which skaters with a significant history of concussions and are therefore removed from the ice decrease their exposure risk and in turn, lower their continuous symptom burden. Lastly, in the non-concussive group, concussion may possibly be under-reported or unidentified, which is supported by the mismatch/inconsistency rate of 34.4% between the matched skater and parent survey reports

Finally, pain ignorance/dismissal and delay in seeking medical assistance rates demonstrates a critical public health issue. Almost half of all skater participants (45.5%) documented dismissing head pain after experiencing head impact on the ice, and almost two-thirds of athletes with concussion histories had partaken in this action. The qualitative feedback implies that the figure skating norms or standards lean toward normalizing and downplaying injury and encouraging skaters to come back quickly or promptly and train again, which are patterns that have already been reported in ice hockey, soccer, and football. Parents also exhibited hesitation and evasion of seeking medical help, as almost two-thirds of parent respondents reported never looking for any official examination even after noticing neurological changes in their child. The results denote the oversight of both concussion education and the response guidelines following events of impact in the figure skating world.

4.2 Pathophysiologic Effects: Inflammasome Activation and Tau Protein Pathways

The neurological symptom patterns noticed in the sample, specifically extensive headaches, fluctuations or changes in mood, decline in overall processing speed and cognition, and disorientation after rotational movements, generally matches the post-concussion symptom groupings depicted in the wide-ranging literature concerning TBI. One possible biological mechanism for these symptoms is the connection between constant subconcussive impacts and neuroinflammatory responses or cascades.

Developing data deriving from human autopsy research and animal models involves the NLRP3 inflammasome, which is a multiprotein innate immune signaling system, as a main enhancer of neuroinflammation after head impact or brain injury. Caspase-1 cleavage and top-bottom leaking of interleukin-1β (IL-1β) and IL-18 follow the activation of NLRP3, which supports an inflammatory local environment specifically in microglia and astrocytes (Heneka et

al.; Gustin et al.). It has been hypothesized that the chronic triggering of this pathway, which can be foreseen from constant subconcussive impacts on the ice in figure skating, expedites the process of tau protein phosphorylation, defective tau removal, and eventually the creation of neurofibrillary tangles (NFTs) congruent with the pathology of chronic traumatic encephalopathy (CTE) (McKee et al.; Stancu et al.).

Although the current study is not meant to measure the biomarkers of inflammasomes or exact levels of tau protein, the pattern of symptoms monitored, specifically the presence of neurocognitive symptoms, such as difficulty with concentration and memory, as well as fluctuations in mood (instability) in athletes without any formal diagnosis or history of concussions, match the premature subclinical neurologic shifts that could show hidden processes of neuroinflammation. Longitudinal biomarker research studies incorporating cerebral spinal fluid (CSF) tau or serum, glial fibrillary acidic protein (GFAP), and phosphorylated tau (p-tau) evaluations in non-retired figure skaters would act as a crucial next systematic phase.

4.3 The Role of Post-Rotation Disorientation

Disorientation following rotational movements was documented at a mean of 1.45/4 among all skater participants and was, strikingly, slightly elevated in the group of non-concussed skaters (mean of 1.56) as opposed to the concussed skaters group (mean of 1.26). The symptom is demonstrative of a distinctive exposure specific to figure skating that was not reported in the majority of sports concussion literature works. Competitive figure skaters perform hundreds of rotations every session of training, which includes spins and twizzles consisting of 2-6+ rotations per element, jumps consisting of multiple (2-4) revolutions, and rotational throws and lifts in the pairs and ice dance categories. The vestibulocerebellum region of the brain goes through constant adaptation that accumulates after every experience of these forces. However, qualitative accounts in this sample propose that the threshold for adaptation could be surpassed in most skaters, specifically towards the ending points or phases of training sessions.

Changes in the function of the otolith, cerebellar floccular habituation, and semicircular canal acclimatization have been associated with the result of disorientation following movement and loading patterns of spinning. In terms of frequent mild subconcussive tension or TBI, the disturbance of these vestibular adaptive processes may constantly recur, conducing to a repeated disproportion in visual, vestibular, and proprioception sensory processing that reveals itself as changes in balance, dizziness, and disorientation that several participants have explained to have progressed over time in their years of skating.

4.4 Limitations

There are multiple limitations in this study that need to be recognized. The first being that the self-report cross-sectional outline is liable to social desirability bias, recall bias, and possible understatement of private medical information, and it prevents causal analysis or logic. The second limitation was that the survey was spread through different skating connections or networks, which proposes selection bias aimed at more active skaters and parents, therefore

possibly over-sampled participants who experienced distressing events. Thirdly, the age range is wide (6-54 years) and the competitive level of the sample is heterogeneous or uneven, which limit the pertinence of synthesized evidence to particular subgroups, and studies in the future should compare cohort-specific differences across distinct age group brackets and competitive levels.

Moreover, the lack of neuroimaging, neuropsychological tests, or biomarker information implies that symptom self-reports are not able to be verified due to insufficient neurological evidence to support the claim. The fifth limitation is that the small amount of matched pairs (n = 32) limits the precision of the cross-informant comparison analyses. Studies in the future should employ prospective cohort designs that implement biomarker assessment, validated tools to evaluate concussions (SCAT6, ImPACT), and baseline calculations pre-season in order to assess neurological differences or changes over time in longitudinal studies.

4.5 Implications for Clinical Practice and Regulation

This study's findings have multiple practical implications. The first is that the figure skating administration, like the International Skating Union (ISU) and U.S. Figure Skating, should formulate and enforce concussion regulations that are sport-specific and that address the repetitive head contact/impact exposure distinctive of figure skating, high proportion of female and pediatric athletes, and the reported norm of minimizing pain. To add on, emphasis on training to identify neurological symptoms linked to frequent head impacts and standardize the practice of seeking medical assistance in coaching education may help confront the symptom-dismissal behaviors found in this study.

In addition, longitudinal neurological surveillance, through the prospective utilization of validated neuropsychological testing softwares digitally, should be incorporated for competitive figure skating athletes, specifically those who have high training load and repetitive falls. Fourthly, further research on vestibular-protective exercises and regulations to strengthen the neck is needed, especially since there was minimal use of neck stability exercise/training (20.2%) in the sample, even with increasing evidence suggesting it can mitigate head kinematics amid impact.

5. Conclusion

This study gives unique cross-sectional data revealing that there is a measurable neurological symptom burden that figure skating athletes experience, which may be linked to constant head impact risk, even without having to receive a formal concussion diagnosis. The presence of pain-dismissing behaviors, mismatch or inconsistency between the matched skater and parent survey reports, and limited employment of medical assistance signifies that possible head impact exposure and neurological symptoms may not be fully acknowledged and or not handled sufficiently in figure skating.

A particularly significant finding of this study is that increased neurological symptom profile/burden is not limited to just the skaters with diagnosed concussion histories. This finding

corresponds to the hypothesis that repetitive subconcussive tension/loading, specifically from forces during landing, rotating mid-air, and spinning, and fall impacts, may trigger neuroinflammatory pathways, which includes the activation of the NLRP3 inflammasome. It may be possible that, even without the presence of a traditionally diagnosable concussion, these pathway processes aid in medically important neurological symptoms.

These results lay the groundwork for prospective future studies incorporating biomarker assessment/sampling, neuroimaging, and longitudinal evaluation to further examine the connection between accumulative biomechanical load or risk factors in figure skating and premature indicators of neuroinflammation and tau protein pathology.

Acknowledgements

The author acknowledges with appreciation the scientific guidance and mentorship of Juan Pablo de Rivero Vaccari, Ph.D., Department of Neurological Surgery, University of Miami Miller School of Medicine.

Works Cited

Bailes, Julian E., et al. "Role of Subconcussion in Repetitive Mild Traumatic Brain Injury." Journal of Neurosurgery, vol. 119, no. 5, Nov. 2013, pp. 1235–45. PubMed, https://doi.org/10.3171/2013.7.JNS121822.

Dubravcic-Simunjak, Sanda, et al. "The Incidence of Injuries in Elite Junior Figure Skaters." The American Journal of Sports Medicine, vol. 31, no. 4, July 2003, pp. 511–17. DOI.org (Crossref), https://doi.org/10.1177/03635465030310040601.

Gustin, Audrey, et al. "NLRP3 Inflammasome Is Expressed and Functional in Mouse Brain Microglia but Not in Astrocytes." PLOS ONE, vol. 10, no. 6, June 2015, p. e0130624. PLoS Journals, https://doi.org/10.1371/journal.pone.0130624.

Hack, Lukas, et al. "Repetitive Subconcussive Head Impacts in Sports and Their Impact on Brain Anatomy and Function: A Systematic Review." International Journal of Sports Medicine, vol. 45, no. 12, Nov. 2024, pp. 871–83. www.thieme-connect.com, https://doi.org/10.1055/a-2342-3604.

Heneka, Michael T., et al. "Inflammasome Signalling in Brain Function and Neurodegenerative Disease." Nature Reviews Neuroscience, vol. 19, no. 10, Oct. 2018, pp. 610–21. www.nature.com, https://doi.org/10.1038/s41583-018-0055-7.

Kerr, Zachary Y., et al. "Agreement Between Athlete-Recalled and Clinically Documented Concussion Histories in Former Collegiate Athletes." The American Journal of Sports Medicine, vol. 43, no. 3, Mar. 2015, pp. 606–13. DOI.org (Crossref), https://doi.org/10.1177/0363546514562180.

McCrory, Paul, et al. "Consensus Statement on Concussion in Sport-the 5th International Conference on Concussion in Sport Held in Berlin, October 2016." British Journal of Sports Medicine, vol. 51, no. 11, June 2017, pp. 838–47. PubMed, https://doi.org/10.1136/bjsports-2017-097699.

McKee, Ann C., et al. "The Spectrum of Disease in Chronic Traumatic Encephalopathy." Brain: A Journal of Neurology, vol. 136, no. Pt 1, Jan. 2013, pp. 43–64. PubMed, https://doi.org/10.1093/brain/aws307.

Montenigro, Philip H., et al. "Clinical Features of Repetitive Traumatic Brain Injury and Chronic Traumatic Encephalopathy." Brain Pathology, vol. 25, no. 3, May 2015, pp. 304–17. DOI.org (Crossref), https://doi.org/10.1111/bpa.12250.

Novak, Zuzana, et al. "Association of Persistent Postconcussion Symptoms with Pediatric Quality of Life." JAMA Pediatrics, vol. 170, no. 12, Dec. 2016, p. e162900, https://doi.org/10.1001/jamapediatrics.2016.2900.

Stancu, Ilie Cosmin, et al. "The NLRP3 Inflammasome Modulates Tau Pathology and Neurodegeneration in a Tauopathy Model." Glia, vol. 70, no. 6, June 2022, pp. 1117–32. PubMed, https://doi.org/10.1002/glia.24160.

The Impact of Quantum Computing on Classical Cryptography: Breaking RSA and ECC with Shor's and Grover's Algorithms By Sanjith Kota

Abstract

For thousands of years, encryption has existed, dating back to ancient Rome with the Caesar Cipher. As the years have progressed, mankind has become even more advanced and has developed the quantum computer, whose principles combine quantum physics and computer science. Recent research displays how quantum entanglement and superposition break modern encryption methods. We will investigate how quantum entanglement and superposition break modern encryption methods such as RSA and Elliptic Curve Cryptography using Shor's and Grover's algorithms. Within these, we can see the use of Hadamard gates, modular arithmetic, and the quantum Fourier transform, only for Shor's algorithm. Hadamard gates play a crucial role in creating superposition, as these gates allow qubits to exist in a perfect mixed state of |0> and |1>, which is critical for both algorithms. Furthermore, modular arithmetic acts as a tool for Grover's and Shor's algorithms, primarily for the factoring of a product of two large prime numbers. Finally, the Quantum Fourier Transform (QFT) has an important role in Shor's algorithm when looking for periodicity, as QFT allows the computer to solve for and find the factors of the product of two large prime numbers. These quantum concepts are all key information for the future, as mankind is slowly advancing beyond this realm of modern, easily breakable encryption methods. With Artificial Intelligence advancing, these modern methods of encryption and storing data will be futile; we must advance towards these new quantum methods, like post-quantum cryptography and quantum key distribution.

Keywords Quantum Superposition, Quantum Entanglement, Shor's Algorithm, Grover's Algorithm, Quantum Cryptography, Quantum Computing, Quantum Key Distribution.

Introduction

In the past, many encryption methods have proliferated: the Atbash cypher from Hebrew, the Caesar cypher from Rome, and the Vigenère cypher from Italy. These cyphers differ from one another but serve the same purpose: they encode messages by reordering the alphabet. For example, the Atbash cypher is like the Caesar cypher, but instead each letter is mapped to its reverse counterpart (Shashikala et al.). Additionally, the Caesar cypher is a symmetric encryption scheme, in which the same fixed number of positions is shifted for each letter. While the Vigenere cypher uses a table called the Vigenere table and acts like a Caesar cypher, this cypher allows you to use more keys instead of just one. This, in turn, would make the cypher stronger (Rathidevi et al.). The Vigenere cypher, in particular, is the most powerful of these three because it took over 300 years to crack it (Rathidevi et al). In this paper, we will dive into modern encryption methods and how they break down with quantum computing.

Over time, encryption methods have become extremely advanced and no longer use this single-key mechanism like the Caesar cypher and the Atbash cypher. The first new type of

encryption method we will discuss is RSA encryption, or Rivest-Shamir-Adleman encryption, named after a computer scientist, a mathematician, and another mathematician. This encryption method relies heavily on “factoring huge numbers” (Sahu and Mazumdar). Additionally, RSA is a really popular public-key encryption method, and specifically, the keys are usually 2048-4096 bits long (Sahu and Mazumdar). Next, Elliptic Curve Cryptography (ECC), as its name suggests, uses elliptic curves to provide security. This algorithm provides the same level of security as RSA encryption, but ECC does it with a smaller key size, making it more efficient in turn (Sahu and Mazumdar). Lastly, AES encryption is more advanced than the other two encryption methods and is really efficient in withstanding robust attacks (Sahu and Mazumdar; Sarkar et al.). This is a symmetric-key encryption with known key sizes of 128, 192, and 256 bits (Sahu and Mazumdar; Gorine and Suhaib).

To fully understand quantum cryptography, we must first look towards qubits. Qubits are unique because they behave differently from regular bits; a bit can either be 0 or 1, which is two options, while in this case, a qubit can have 3 options: 0, 1, or both. The “both” option is when the two states are in superposition, and this is extremely unique and useful, as it allows the bit to represent 2^n states at the same time, where n is the number of qubits. Qubits play an integral part in quantum cryptography. We can see this in Figure 1.

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

Figure 1: This is the equation that represents superposition; $|\psi\rangle$ represents the wave function where this qubit represents the combined states of $|0\rangle$ and $|1\rangle$. α and β in this equation are complex numbers that represent the probability of the qubit being in state $|0\rangle$ or $|1\rangle$. The above equation works with the requirement that $|\alpha|^2 + |\beta|^2 = 1$ (Sahu and Mazumdar).

Additionally, another crucial part of quantum cryptography is understanding superposition and the role it plays in creating qubits, which in turn is important for Shor’s and Grover’s algorithms (Mishra, Agarwal & Patel; Portmann and Renner; SAFNA). Quantum superposition states that a quantum system can have multiple states existing simultaneously. As we can see, having multiple states exist simultaneously creates qubits, which are really useful for quantum cryptography. An example of superposition can be seen with the double-slit experiment, where electrons, when they pass through multiple slits, create an interference pattern. This pattern indicates that the electrons act as if they have passed through both slits, which is proof of superposition.

Entanglement plays an important role in quantum cryptography. When two quantum systems become interlocked with each other, the quantum state of one of the particles directly determines the state of the other one. Basically, there is a correlation between two things (Shannon et al.).

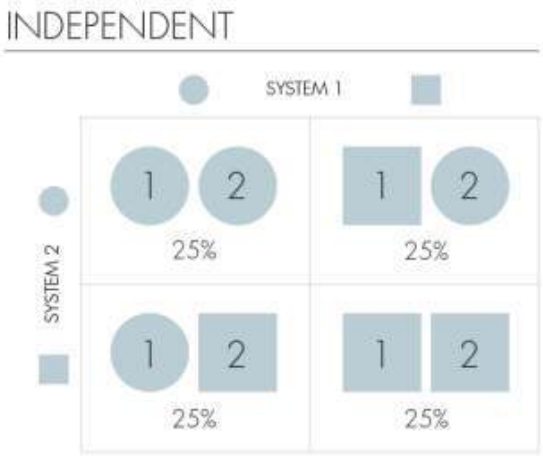


Figure 2: 2 non-entangled systems, showing that probability is equal (Figure 1 in Shannon et al.).

Figure 3: 2 entangled systems, showing that probability is not equal (Figure 2 in Shannon et al.).

In the two images, we can see that Figure 2 shows two independent systems, with equal probabilities of 25% each, because they are independent, meaning one system doesn't affect the other. However, in Figure 3, we can see that two entangled systems are correlated. Likewise, to know if a system is truly entangled, correlation is not sufficient (Shannon et al.). Entanglement has a lot of uses in encryption; if one person has one of the photons in a pair and another person has the other, they could be entangled and give your message more security against any eavesdroppers. If an eavesdropper were to try to intercept your message, you would immediately know, as quantum decoherence states that a subtle change in the environment can affect the quantum state of a system. One would immediately know if an eavesdropper is trying to intercept a message, as there would be a change in one of the photons' quantum states, and since the two are entangled, this would affect the other photon too.

In conclusion, in order to fully understand the basis of quantum cryptography with Shor's and Grover's algorithms, these concepts need to be taken as a basis beforehand. Firstly, RSA and ECC display how Shor's and Grover's algorithms can break these two modern and advanced encryption techniques. Secondly, qubits play an important part in breaking these modern encryption methods, as does superposition. Lastly, entanglement is another key factor in quantum cryptography because of its ability to immediately change the quantum states of two interlocked particles. All of these factors play an important role in Grover's and Shor's algorithms, which will be touched on later.

Quantum Cryptography plays an important role in society. Quantum cryptography rapidly advances due to researchers looking for ways to secure communications against increasingly powerful cyber threats. One new development that is occurring is Quantum Key Distribution (QKD), which utilizes quantum mechanics to allow two different parties to share encryption keys with a very high level of security (Scarani et al.). Similarly, the encryption of quantum data would be an important outlook in the future (Alagic et al.; Mogilevskaia). At the same time,

Artificial Intelligence (AI) plays a very important role in improving quantum cryptography; as more AIs advance and improve, quantum cryptography will only grow stronger and stronger. This is why we need to learn more about quantum computing, as its benefits would help create more secure information, which would then preserve our future.

Literature Review

In this paper, the main method utilized was searching with keywords. When searching for my sources, I used keywords such as my research question, "how quantum entanglement and superposition break modern encryption methods such as RSA and Elliptic Curve Cryptography using Shor's and Grover's algorithms," or keywords such as "quantum encryption AND quantum entanglement". I used connector words such as "AND" in order to provide search results and sources that have the concepts that I was searching for. During my search, I primarily used the database and search engine of Google Scholar because it provided me with the most accurate sources as I went through each of my sources. I was able to see that they were really credible and cited by a multitude of people. To conclude, during my search for sources to prove my thesis, I used Google Scholar as a database, and I used keywords such as my research question and keywords that used connectors within them in order to get the most accurate results.

Entanglement

Quantum entanglement is one of the most fundamental and important phenomena not just in quantum cryptography but in the whole of quantum physics. Entanglement allows quantum systems to perform calculations needed for Shor's and Grover's algorithms, but cannot be accessed by regular computers. When two particles become "entangled" with each other, you cannot physically know the state of one particle without considering the other particle. This entanglement state, in turn, allows quantum computers the ability to process and connect information across many qubits at the same time. Entanglement plays a very important role in Shor's and Grover's algorithms, which solve problems more efficiently than classical computers, like RSA and ECC. In the future, as Shor's and Grover's algorithms improve, cryptographic systems like RSA and ECC will become obsolete, and we can see how entanglement plays a role in these two systems with the EPR Paradox.

The Einstein-Podolsky-Rosen Paradox (EPR Paradox) states that when two qubits are entangled, each of their quantum states cannot be determined independently, as the measurement of one state directly affects the other (Einstein, Podolsky & Rosen; Blaylock). We can see this defined in Figure 4.

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$$

Figure 4: This figure represents the EPR paradox in the Psi state, where $|\psi\rangle$ represents the wave function. $|00\rangle$ and $|11\rangle$ are basis states, and $\frac{1}{\sqrt{2}}$ normalize the states (Horodecki et al.).

$$|\psi\rangle = \frac{1}{\sqrt{2}} (|01\rangle \ + \ |10\rangle)$$

Figure 5: This figure represents the EPR Paradox in the Phi state, where $|\psi\rangle$ represents the wave function. $|00\rangle$ and $|11\rangle$ are basis states, and $\frac{1}{\sqrt{2}}$ normalize the states (Horodecki et al.).

The thought experiment of the EPR Paradox was developed to challenge the principles of quantum mechanics. Einstein, Podolsky, and Rosen argued that if their theory was correct, then the measurement of one particle would directly affect the other particle (Einstein and Podolsky & Rosen). Einstein, Podolsky, and Rosen's finding basically supports the idea of entanglement, and we can see the equation represented above in Figures 4 and 5. Thus, the EPR Paradox serves as a foundation for Shor's and Grover's algorithms to break modern encryption methods. Because of the entangled states across many particles, Shor's and Grover's algorithms can coordinate information very efficiently, which is what allows them the ability to break these modern encryption methods.

In conclusion, entanglement is a key instrument across Shor's and Grover's algorithms because of what we can see in Figure 4 about the EPR Paradox. Entanglement enables coordination and information processing more efficiently than in regular computers, which is why Shor's and Grover's algorithms can easily break modern encryption methods like RSA and ECC.

Hadamard Gates

The Hadamard gate plays a very instrumental role in Shor's and Grover's algorithms. Hadamard Gates play a very useful and important role in Shor's and Grover's algorithms because of their ability to create superpositions of zero and one states (Flarend and Hilborn). These states are very common in that they have their own notation of |+> and |-> (Flarend and Hilborn).

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) = |+\rangle$$

$$H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) = |-\rangle$$

Figure 6: In the figure, we can see the two equations for the Hadamard gate states of 0 and 1 (Flarend and Hilborn).

This Hadamard gate is named for the Hadamard transform, which, in the context of quantum computing, allows the simultaneous application of Hadamard gates to multiple qubits (Flarend and Hilborn). In the context of Shor's and Grover's algorithms, Hadamard gates allow

these algorithms to perform so efficiently because of this ability to apply uniform superposition over all qubits.

Grover's algorithm is a really powerful quantum algorithm that is superior to modern computers, so much so that it can break modern encryption methods like RSA and ECC. This search algorithm is so efficient that it can be solved in $O(\sqrt{n})$ operations; the notation represents Big O notation, which shows the efficiency for certain operations (Moreno). For the quantum search procedure of Grover's algorithm, it starts with the computer being in the $|0i\rangle^{\otimes n}$, where the $\otimes$ represents the tensor product notation (Moreno). After that, we can use the Hadamard gate on each of the qubits so that they reach a superposition state. This would then leave the system in the state in Figure 7, and we can see the system as a quantum circuit within Figure 8 (Moreno).

$$G = H^{\otimes n}(2|00\rangle - I)H^{\otimes n}O = (2|\psi\psi\rangle - I)O$$

Figure 7: In this figure, we can see the definition of the G operator. Where $H^{\otimes n}$ represents each qubit within the system, and where O represents the O operator or oracle operator, and I is the identity operator (Moreno).

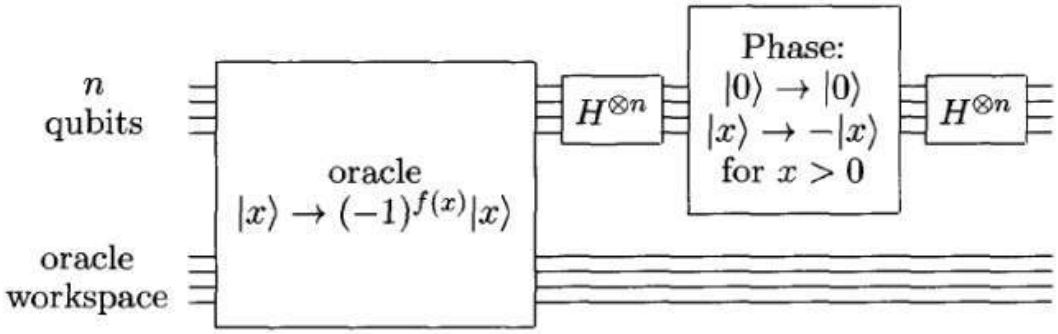


Figure 8: A pictorial representation of the equation above as a quantum circuit for Grover's algorithm (Figure 1 in Moreno).

In both Figure 7 and 8, the oracle operator is represented as O within the equation. Within the context of Grover's algorithm, the oracle operator is used to mark the correct solution of a search problem by flipping its phase. The oracle operator acts as a filter that identifies states meeting the right condition; this in turn leads to a quadratic speedup in efficiency from O(n) to $O(\sqrt{n})$ (Moreno).

In Shor's algorithm, Hadamard gates primarily play a role in period finding and are mainly used with the Quantum Fourier Transform (QFT). Firstly, Hadamard gates provide the same first step in Shor's algorithm as Grover's, as the Hadamard gate is applied to each qubit $|0i\rangle^{\otimes n}$ so that it transforms each of the states into an equal superposition. The Hadamard gate plays a role like this for the quantum computer to evaluate all solutions at the same time. However, where Grover's and Shor's algorithms differ is that Hadamard gates play a role in

QFT, which is only used in Shor's algorithm specifically. The reason that Hadamard gates only appear in Shor's algorithm is to allow the quantum computer to find the period from a quantum state.

In summary, Hadamard gates play a very crucial role in both algorithms. Hadamard gates allow Grover's algorithm to gain the quadratic speedup in efficiency from O(n) to O($\sqrt{n}$), and Hadamard gates play a role in period finding within Shor's algorithm. These reasons allow for Shor's and Grover's algorithms to easily break modern encryption methods like RSA and ECC, because of their incredible efficiency in searching.

Modular Arithmetic

To begin, modular arithmetic is calculations involving the modulus operator. The modulus operator is similar to the division operator; however, when used, it gives the remainder of the two numbers used (Horowitz). The notation of this function to find the result A is given as A = b (mod(m)), where b and m are some integers (Horowitz). Modular arithmetic is crucial in Shor's and Grover's algorithms because it can allow them to factor large integers exponentially faster than regular methods.

In Shor's algorithm, modular arithmetic is an essential part of finding the period of a function. A modular function is given by $f(x) = a^x \bmod(N)$, where a is a base coprime to N. For this function, there is a period defined as r, and it is the smallest possible integer solution for the equation $a^r = \bmod(N)$. We can see that everywhere in Shor's algorithm the modulus operator shows up, displaying that it's a very integral part of Shor's algorithm. Additionally, modular arithmetic is essential because it allows both Shor's and Grover's algorithms to work with remainders of large numbers rather than the numbers themselves, which in turn makes them easier to implement.

In Grover's algorithm, modular arithmetic is mainly used with the oracle operator so that this operator can check every qubit and see if they belong by using modular operations. The reason why modular arithmetic plays a key role in Grover's algorithm is because it keeps the search space finite, as the remainder always stays constant, which in turn can simplify the implementation of the oracle operator. Similar to Shor's algorithm, it allows the problem to be expressed in terms of remainders, which is why modular arithmetic is so important for these two quantum algorithms.

Quantum Fourier Transform

QFT plays a very crucial role in Shor's algorithm, and out of Grover's and Shor's algorithms, it is only local within Shor's, and the reason for this is because of a specialty within Shor's algorithm. In Shor's algorithm, QFT plays a role in period finding, which involves the use of Hadamard gates and QFT. Period finding plays a very crucial role in Shor's algorithm in order to get the desired efficiency that can break RSA and ECC. QFT is a linear transformation of qubits, and it is the quantum version of the well-known Fourier Transform. QFT plays a crucial role in Shor's algorithm because of its factoring of discrete logarithms.

QFT takes the position state |x> to the corresponding momentum state |p> (Weinstein, Lloyd & Cory). This can be seen in Figure 9.

$$\mathrm{QFT_q}|x\rangle \rightarrow \frac{1}{\sqrt{q}} \sum_{p=0}^{q-1} e^{2\pi i ap/q}|p\rangle$$

Figure 9: This represents the equation showing how QFT takes a position state to its corresponding momentum state, where q is the dimension of the system's Hilbert Space, and where a is the input position/state index (Weinstein, Lloyd & Cory).

In general, QFT takes the position input amplitudes and then converts them into the corresponding momentum amplitudes (Weinstein, Lloyd & Cory). This transformation can be seen in Figure 10.

$$\mathrm{QFT_q} \sum_x f(x)|x\rangle \rightarrow \sum_p f(p)|p\rangle$$

$$f(p) = \frac{1}{\sqrt{q}} \sum_a e^{2\pi i ap/q}|p\rangle$$

Figure 10: This is the equation that represents the transformation of the input amplitudes into the momentum amplitudes as seen above (Weinstein, Lloyd & Cory). The second equation within the figure shows the coefficients of f(p) (Weinstein, Lloyd & Cory).

In Shor's algorithm, period finding plays a key role within it. The classical Fourier transform reveals the periodicity within regular functions, while the QFT reveals the periodicity within wave functions, which is how Shor's algorithm operates (Weinstein, Lloyd & Cory). After the period is found, defined by r, we can use this to analyze and factor N. If r is even and $a^{r/2} \neq -1 (\mathrm{mod}\ N)$, then we can see that the factors of N would be $\gcd(a^{r/2} - 1, N)$ and $\gcd(a^{r/2} + 1, N)$.

In summary, QFT plays a key role within Shor's algorithm specifically, as it allows the change in efficiency from exponential time to polynomial time. Normally, N would increase exponentially with the key size for RSA/ECC, but with Shor's algorithm it reduces it to a polynomial increase. This speed-up is the result of QFT's role in period finding, as QFT reveals the period within wave functions (Weinstein, Lloyd & Cory). Once this period is found, we can use it to analyze N and see how N factors.

Conclusion

The development of quantum computing and quantum cryptography poses a great threat to the security of society. As society develops, there will be more advances in quantum cryptography, like Shor's and Grover's algorithms. Throughout this paper, we saw two modern encryption methods, those being RSA encryption and ECC, break under Shor's and Grover's algorithms. Usually, RSA and ECC would be secure against classical computers and pose a strong defense against any cyberattacks, but quantum computing has shown that this security would eventually fail. Both Grover's and Shor's algorithms can solve problems efficiently within polynomial time. As a result of this, the security behind RSA and ECC would no longer prove effective against these two algorithms.

However, as these two algorithms advance, new technologies are emerging to counter Shor's and Grover's algorithms. One promising new development is Quantum Key Distribution (QKD), which provides a theoretically perfect amount of security by using the principles of quantum mechanics to exchange the encryption keys. The reason for this high-level of security is that any attempt to see the encryption key would immediately disturb it, due to the entanglement between superposition states. QKD offers a tremendous amount of security if someone is trying to eavesdrop, and it does not rely on mathematical concepts. However, QKD faces a lot of limitations, such as infrastructure costs, transmission distance, and scalability.

Additionally, as society advances, AI is also supposed to take a leap. In the future, I see that AI could assist with post-quantum cryptography, as AI can assist researchers with developing new cryptographic algorithms. As we see organizations turn away from modern encryption methods, such as RSA and ECC, AI-powered security tools may help with key management, monitor anomalies, and promote the development and adoption of quantum-resistant technologies. However, AI can also pose serious risks and threats to this system. With AI, attackers can use machine learning to improve phishing campaigns and develop more advanced cyberattacks. As a result, AI would likely play a role on both the offensive side and defensive side in the future of technology.

Ultimately, the rise of quantum computing starts a new era of security, and not the end of the old one. In the future, protecting digital information would require a combination of post-quantum cryptographic algorithms, emerging technologies like QKD, and Artificial Intelligence. Companies that start preparing today will be better positioned in the future to maintain their privacy and confidentiality as these quantum algorithms continue to evolve.

Acknowledgements

I would like to give my sincerest gratitude to my mentor throughout this project, Professor Ryan Farber. He was a prominent figure in helping me understand the content before writing this paper. Without him, I probably would have found myself looking through the depths to find lectures and courses on this material. Additionally, he helped me with getting the papers needed; even if I did not have access to them, he was able to find the papers and give them to me so that I could use them on my project. Finally, any doubt that I had on this project or anything that could have been improved, Professor Ryan Farber often provided insightful feedback.

Works Cited

Alagic, Gorjan, et al. "Computational security of quantum encryption." International Conference on Information Theoretic Security. Cham: Springer International Publishing, 2016.

Blaylock, Guy. “The EPR Paradox, Bell’s Inequality, and the Question of Locality.” American Journal of Physics, vol. 78, no. 1, 2010, pp. 111–20.

Flarend, Alice, and Robert Hilborn. Quantum computing: from Alice to Bob. Oxford University Press, 2022.

Einstein, Albert, Boris Podolsky, and Nathan Rosen. "Can quantum-mechanical description of physical reality be considered complete?." Physical review 47.10 (1935): 777.

Gorine, Adam, and Muhammad Suhaib. “Exploring AES Encryption Implementation through Quantum Computing Techniques.” Am. J. Comput. Sci. Technol., vol. 7, 2024, pp. 139–55.

Horodecki, Ryszard, et al. “Quantum Entanglement.” Reviews of Modern Physics, vol. 81, no. 2, 2009, pp. 865–942.

Horowitz, Ellis. "Modular arithmetic and finite field theory: A tutorial." Proceedings of the second ACM symposium on Symbolic and algebraic manipulation. 1971.

Mishra, Shreshtha, Ojusav Agarwal, and Sunil Kumar Patel. "Quantum decryption using shor’s algorithm." 2024 IEEE Pune Section International Conference (PuneCon). IEEE, 2024.

Mogilevskaia, Iliana. Quantum Cryptography as an Alternative to Modern Cryptography. 2019.

Moreno, Gerardo García. "Grover’s Algorithm." Universidad Complutense de Madrid (2018).

Portmann, Christopher, and Renato Renner. “Security in Quantum Cryptography.” Reviews of Modern Physics, vol. 94, no. 2, 2022, p. 025008.

Rathidevi, M., R. Yaminipriya and S. V. Sudha, "Trends of cryptography stepping from ancient to modern," 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), Coimbatore, India, 2017, pp. 1-9.

SAFNA, CS. QUANTUM COMPUTING IN CRYPTOGRAPHY. 2025.

Sahu, Swastik Kumar, and Kaushik Mazumdar. “State-of-the-Art Analysis of Quantum Cryptography: Applications and Future Prospects.” Frontiers in Physics, vol. 12, 2024, p. 1456491.

Sarkar, Bikramjit, et al. “A Survey on the Advanced Encryption Standard (AES): A Pillar of Modern Cryptography.” International Journal of Computer Science and Mobile Computing, vol. 13, no. 4, 2024, pp. 68–87.

Scarani, Valerio, et al. "The security of practical quantum key distribution." Reviews of modern physics 81.3 (2009): 1301-1350.

Shannon, Keith, et al. “On the Use of Quantum Entanglement in Secure Communications: A Survey.” arXiv Preprint arXiv:2003.07907, 2020.

Shashikala, K. S., et al. "Cryptography and Blockchain: Building Blocks of Secure Decentralisation." Applying Blockchain Technology. Auerbach Publications, 2025. 155-168.

Weinstein, Yaakov S., et al. "Implementation of the quantum Fourier transform." Physical review letters 86.9 (2001): 1889.

How Plants Can Be Grown In Space By Krishi Boppana

Introduction

"Recent space exploration has extended the field of plant biology, allowing for future studies on life support farming on distant planets"(1). When astronauts are travelling in space they will need the proper nutrition like they have back home and while some people might argue that the freeze dried option can be beneficial for the astronauts they just do not have enough nutrients so what they can do is grow plants in their shuttle. Growing plants in space requires special conditions and nutrients along with water.

Background information

During long space missions, growing plants will provide nutrients necessary for astronauts to survive. Even if plants can adjust to microgravity plants still require special growth conditions so this is where scientists come into the frame. Not only do plants provide nutrition for humans but scientists can also study how they react to the harsh conditions. Plants on earth endure harsh conditions like for example cacti survive in deserts by growing long roots to far water sources and plants also survive underwater in low-light conditions by using chemical energy emitted by hydrothermal vents."Plants are investigated under red- and blue-light irradiation at microgravity settings of 0.1 *g*, 0.3 *g*, and 1.0 *g*, modelling environmental conditions on the moon (0.17 *g*) or Mars (0.38 *g*), respectively [12]. This study shows that directional light could act as a signal to govern directional plant growth even in the case of long-term low-gravity conditions"(1). This study also shows that there are molecular gravity sensing properties to direct the plant's growth to favorable conditions. During a mission, scarce resources like power, crew involvement and data must be minimized to support the mission. "The current spaceflight facilities vary in complexity from simple Petri dishes to closed-loop feedback-controlled chambers that regulate biologically relevant parameters such as photosynthetic illumination intensity and quality, diurnal cycle, temperature, relative humidity, moisture, atmospheric constituency, and even fractional gravity"(4). In spaceflight, these are the conditions that scientists are studying.

Research Findings

One of the main reasons why scientists grow plants in space is to test their reactions to different conditions. Testing the plants' reactions to different conditions is very beneficial to astronauts because it could determine whether the plant can resist the conditions. "Plant experiments have been performed in space since the launch of Soviet/Russian Salyut 1, the first space station [78]" (Plants in Microgravity- NLM). Scientists had been testing plants in space for decades and now they have even successfully grown many vegetables, flowers and fruits. Without the force of gravity many factors that determine the plants' growth have been altered. "E.g., studies with the Micro-Tom dwarf tomato variety highlighted the negative effect of microgravity on nutrient acquisition and plant growth, including the overall yield [112]"(1). This

explains that gravity can really affect plant growth for certain types of plants. Plants have a way of adapting to space flight conditions and it is by using a type of undiscovered stress related genes. “An experiment involved the *A. thaliana* seeds germinating and growing under simulated gravity similar to that of the moon’s (0.17 *g*) and Mars’ (0.38 *g*) gravity. The results of these partial gravity simulation experiments displayed an increment in cell proliferation and decreased cellular growth under the levels of moon gravity [133]”(1). That explains in lower gravity conditions the plants can not function as well as in regular gravity. In NASA they have a growth chamber called Advanced Plant Habitat(APH) where they use LED lights for the plants energy and a clay substrate to deliver water. There are over 180 sensors used to view its water recovery, atmospheric content, moisture conditions levels and temperature. After the plant is fully grown scientists collect preserved samples then send them back to earth to be tested on how space and gravity impacted the samples.

Real world Examples

Plants respond to the environment that they grow in like for example on earth the plant shoots up and roots go down but in space where the gravity is little to none the roots grow closer to water sources and the plant grows toward the light (3). To further benefit the plant’s growth the plants will be placed in a type of soil that will cycle around the nutrients and water. “In the absence of gravity, plants use other environmental factors, such as light, to orient and guide growth”(2). Plants will try to find out many ways to get nutrients if one is unavailable. All of these factors are being tested by NASA’s Vegetable Production System (Veggie), where they study plant growth in microgravity conditions and to provide astronauts with fresh fruits and vegetables. Three types of lettuce have been grown by Veggie where these examples include: Chinese cabbage, mizuna mustard and red Russian kale. “Foods like berries, certain beans and other antioxidant-rich foods would have the added benefit of providing some space radiation protection for crew members who eat them”(2). Some foods can provide astronauts with health benefits and vitamins.

Conclusion

Ultimately, scientists on the International Space Station(ISS) are currently researching how the conditions in space are affecting the growth of plants in space. Viewing how plants are affected in conditions such as microgravity or higher radiation levels is very important to measure because during long space travel, astronauts will need proper nutrition from fresh fruits and vegetables. An example of plants being tested is in NASA they have an operation called “Veggie” where they study plant growth in conditions of microgravity so they can grow fresh food for astronauts and they have grown a few types of plants like chinese cabbage, mizuna mustard and red Russian kale. In microgravity plant growth seems to decrease slightly because gravity is an indicator for plants where to grow. Knowing how plants can be grown in space will be the future for long space missions and even living on a different planet such as Mars.

Works Cited

1. Baba, Abu Imran, et al. “Plants in Microgravity: Molecular and Technological Perspectives.” International Journal of Molecular Sciences, vol. 23, no. 18, 11 Sept. 2022, p. 10548, https://doi.org/10.3390/ijms231810548.
2. NASA. “Growing Plants in Space - NASA.” NASA, NASA, 9 June 2023, www.nasa.gov/exploration-research-and-technology/growing-plants-in-space/
3. “Plants in Space.” CALS Magazine, 24 June 2022, magazine.cals.ncsu.edu/plants-in-space/.
4. Reed, David W, and Carme Bosch. “Engineering Perspectives of Growing Plants in Space.” Journal of the Indian Institute of Sciences, vol. 103, no. 3, 18 May 2023, pp. 797–805, https://doi.org/10.1007/s41745-023-00369-6.

ChatGPT and Social Etiquette By Nicole Wang

Abstract

With hundreds of millions reportedly actively using ChatGPT, it behooves us to study how our interactions with this platform impact our brain and spill over into other areas of our lives. Does how we interact with a Chatbot influence how we communicate with others? Do our brains distinguish between talking to ChatGPT versus another human being? If these lines are blurred, as more recent research suggests, then does how we interact with Chat have implications for our real-world relationships? These kinds of questions are what have prompted a discussion of whether people should be encouraged to be polite to chat bots because repeated interactions on these platforms have been shown to bleed over into daily interactions. This paper argues that it is indeed important to be polite to ChatGPT because the way we interact here does indeed impact how we treat others.

Introduction

"Tens of millions of dollars well spent," Sam Altman, the CEO of OpenAI, responded when asked about the cost OpenAI sustains from users typing pleasantries into ChatGPT (Altman). Key to his response is the phrase, "well spent," which implies a return. So, the real question becomes: a return for whom?

The instinctive answer is that nobody benefits. ChatGPT is a system that processes linguistic form "without any reference to meaning: a stochastic parrot," not a conscious entity (Bender et al. 616). It has no ability to interpret meaning, no capacity for suffering, and therefore no right to moral consideration. From this view, saying "please" to a language model is a paradox, and Altman's "well spent" is sentimental nonsense.

But this conclusion assumes the only justification for politeness is the recipient's ability to receive it, and ignores the impact on the person practicing it. Politeness is not only transactional; it is a practice that solidifies the behavioral and cognitive habits of the person performing it. The relevant issue is not whether ChatGPT registers the gesture, but whether the person making it does.

This essay argues that people should be polite to ChatGPT because of the ramifications for humanity. To support this statement, it is critical to first, establish the ubiquity of ChatGPT and thus its ability to alter human interaction at scale; second, understand how ChatGPT was developed and what it is; third, detail the neurological impact of repeated interaction on the human brain; and fourth, explore social ramifications.

Development of ChatGPT

In terms of presence, ChatGPT's monthly active users grew continuously throughout 2025, with cumulative mobile users climbing sevenfold from 2023 to hundreds of millions by early 2026 (We Are Social et al.). Yet scale alone does not explain the risk. OpenAI's own analysis found that a 60% majority of interactions are deeply personal, with 49% "Asking,"

treating ChatGPT as an advisor, and 11% “Expressing,” characterized by personal reflection and play. Fewer than half are purely task oriented. Even by OpenAI’s own findings, the dominant form of engagement is social, not transactional (“How People”).

Given ChatGPT's widespread presence and use, it is necessary to understand the platform itself. The architecture underlying ChatGPT was not designed for conversation or moral reasoning. The Transformer was built for machine translation between languages (Vaswani et al.). Nothing social or moral was written into its underlying mechanism. ChatGPT's conversational voice is a layer added through subsequent training, designed to imitate human interaction.

Achieving that replication of conversational patterns required a scale expansion. GPT-3, trained on 175 billion parameters and 300 billion tokens of internet text, learned to mimic human communication statistically, through the absorption of the politeness, warmth, and social conventions embedded in its training data, without understanding any of it (Brown et al. 8). A language model is “a system for haphazardly stitching together sequences of linguistic forms…without any reference to meaning” (Bender et al. 616). What feels like understanding is filtered by our “predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do” (Bender et al. 616).

The conversational warmth users encounter was further honed through Reinforcement Learning from Human Feedback (RLHF), a technique in which human evaluators rank outputs the model provides according to their preferences; those preferences are then used as a reward signal to fine-tune the model (OpenAI, “Aligning”). RLHF was developed precisely because certain behavioral goals could not be written explicitly as code, as they required human judgment to actualize (Christiano et al.).

As OpenAI used human feedback to train ChatGPT to mimic a sentient being, even if its responses do not signal understanding, in its 2022 announcement, it also described ChatGPT as a model that can “admit its mistakes, challenge incorrect premises, and reject inappropriate requests,” characteristics that attribute moral agency to what is, structurally, a pattern-reproducing algorithm (OpenAI, “Introducing”). The social facade of ChatGPT is thus the product of two-part encoding, reinforced by OpenAI messaging. Both of these processes lead users to encounter something that looks and sounds like a real conversational individual.

Moreover, the human tendency to anthropomorphize machines is well-documented and predates ChatGPT. Experiments conducted in the late 1990s by Reeves and Nass demonstrated that politeness towards machines is “automatic, unavoidable, and happens more often than” people realize (Reeves and Nass). Indeed, Joseph Weizenbaum, creator of the ELIZA chatbot in 1966, was shocked to observe that his own secretary became emotionally absorbed in conversations with what he knew to be a simple computer program, asking him to leave the room so she could speak with it privately (Weizenbaum).

Implications for the Human Brain

Given ChatGPT's ubiquity and the human tendency to anthropomorphize machines, the case for politeness rests ultimately on what repeated interaction does to the human brain. Habits are "relatively automatic, and unconscious…inflexible and rigid," consisting of "an ordered, structured action sequence that is prone to being elicited by a particular context or stimulus" (Graybiel 361). Crucially, habit formation is not restricted to motor sequences; it extends to cognitive routines, manifesting as habits of thought. Within the basal ganglia, repeated sequences are "chunked" into single executable units, transforming deliberate behavior into reflex (Graybiel 361). A user who is rude to ChatGPT fifty times in a day is not performing fifty discrete acts; instead, they are reinforcing a chunked pattern of incivility that the brain will retrieve automatically the next time a similar context arises.

This pattern is the foundation of cognition formed through repetition. The brain, which does not stop rewiring itself in adulthood, continuously reshapes itself in response to repeated and frequent actions. Daily behavior ultimately shapes its lasting structure. Nicholas Carr asserts his findings on neuroplasticity in The Shallows on the core principle, "We become, neurologically, what we think" (29). He argues that intellectual technology changes what and how we think, because any repeated experience influences our synapses. The tools one engages with the most intensively, over time, become the architecture of the mind. Notably, AI interaction is the newest and most frequent instance of this older phenomenon. Conversing with ChatGPT repeatedly daily initiates a sustained training of the brain, habituating oneself to the conversational style adopted toward the machine.

What makes this consequential is that the changes are not easily undone. Carr establishes that plasticity is not the same as elasticity: our neural loops "don't snap back to their former state the way a rubber band does; they hold onto their changed state" (30). This is critical in demonstrating that it is the user who registers every exchange, and that the brain reverts to the default continuously built through the rehearsal of thousands of interactions with the machine.

Context Blurring

The brain's inability to separate one context from another is what turns this from an internal conflict into a genuine risk. The ease with which humans anthropomorphize machines supports the neurological record, which suggests this separation is far more difficult to maintain than it appears. Once a behavior has been chunked, control over it moves from the prefrontal cortex, controlling deliberation, to the dorsolateral striatum, driving automaticity, leaving the behavior "resistant to change" even when the actor consciously desires otherwise (Smith and Graybiel). Knowing a habit is undesirable does not deactivate it. The brain has no reliable mechanism to keep the manner practiced with machines separate from the manner used with people.

Even more crucial is that this susceptibility is not evenly distributed. A lack of social connection has been identified as one of the strongest catalysts of anthropomorphism, which means the lonelier a person is, the more likely they are to relate to the system as though it were conscious and feeling (Bartz et al.). A joint study by the MIT Media Lab and OpenAI, drawing

on roughly forty million interactions and a four-week trial of about a thousand participants, found that higher daily usage "correlated with higher loneliness, dependence, and problematic use, and lower socialization," and that those who bonded most closely with the system were the likeliest to be lonely and to rely on it even further (Fang et al.). The people most exposed to the habits formed in these exchanges are the people who are the least capable of protecting themselves against them.

A study of long-term Replika users documents genuine bonds, including descriptions of love and friendship, formed by people who remained fully aware that their companion was not conscious (Skjuve et al.). Other research found users describing the AI as a "best friend" that is "always there," defending the relationship's value while acknowledging its asymmetry, and identified human-AI friendship as a new relational category (Brandtzaeg et al.). Yet there are harsh consequences: users have reported experiencing distress when software updates altered a chatbot's "personality," and in some cases withdrawal from human relationships in favor of the more accommodating machine (Laestadius et al.). Experimental work confirms the reverse is also true, showing that tailored and empathetic AI responses produce stronger emotional validation (Merrill et al.).

The danger in this bond was named more than a decade ago, in warning that "digital connections and sociable robots may offer the illusion of companionship without the demands of friendship" (Turkle). Human relationships require reciprocity, patience, and a tolerance for argument, yet a relationship manufactured around obedience and efficiency requires none of these traits. Because the system carries no reciprocal obligation, the user is left as the only party capable of supplying any, and politeness therefore becomes one of the few remaining forms of self-discipline available within such an imbalanced exchange. Hence, an even larger social risk follows: if people grow accustomed to an interlocutor that never demands and always obeys, they may treat genuine human conversation, which requires mutual effort, as an overly complicated version of the effortless conversation the machine provides.

Social Implications

This is where a private problem becomes a public one. Once this inability to compartmentalize is multiplied across hundreds of millions of users, individual routines become public norms. What follows is that neural pathways built through the most frequent exchange then determine interpersonal interactions — moving outwards from first the family, then the workplace, and finally civic order.

The innermost of these circles is family, and evidence of that leakage already exists. A four-month study of households using the Amazon Echo found that the device "definitely changes how people converse," with mothers reporting that their children began demanding things in command-style speech: "Mommy, bring me orange juice" (qtd. in "Alexa!"). The command-like speech developed through interactions, without conscious decision, influenced how children addressed their parents. A study argues that incivility spirals, one rude act provokes another across relationships (Andersson and Pearson).

The next circle is the workplace, where a person's conversational defaults dictate how they treat individuals they have no personal incentive to accommodate. A study of 440 employee-leader pairs found that incivility carried in from outside the office positively predicts incivility committed at work, with diminished self-control transmitting the habit from one domain into the other (Deng et al.). This mechanism is crucial because self-control is precisely what solely compliant exchanges fail to cultivate. This migration also works in both directions, as workplace rudeness has been shown to cross back into the home, eroding marital and family satisfaction, so the two domains create a self-reinforcing loop (Ferguson). A workforce composed of individuals whose manners of communication have been molded by thousands of frictionless, obligation-free interactions does not just emphasize increased rudeness; it also leads to the loss of the reciprocity that allows a group of strangers to cooperate at all.

Finally, the outermost circle is civic order. Social capital is the cumulative product of interpersonal exchanges that serve as a prerequisite for civic engagement and democratic governance (Putnam). If a society's ability to self-govern rests on the developed trust between its members, then whatever erodes that trust erodes that capacity. By spring 2024, only 22 percent of Americans said they trust the federal government to do the right thing just about always or most of the time, down from 77 percent in 1964 (Pew Research Center). This is an active process, and ChatGPT enters it at a scale no prior technology has. Essentially, distrust of others is associated with the inability to engage in civil communication about shared problems, which is what self-governance requires. A democracy cannot function if its members are incapable of treating one another as people worth reasoning with.

Conclusion

Altman's “well spent” now rings with clarity — the price safeguards not a product, but our capacity to treat one another with respect — in a world where many of our conversations may no longer be with a person at all.

Works Cited

"Alexa! Why Are My Kids Yelling at You?" BYU Marriott Magazine, Jan. 2019.

Altman, Sam [@sama]. "tens of millions of dollars well spent--you never know." X, 16 Apr. 2025, x.com/sama/status/1912646035979239430.

Andersson, Lynne M., and Christine M. Pearson. "Tit for Tat? The Spiraling Effect of Incivility in the Workplace." Academy of Management Review, vol. 24, no. 3, 1999, pp. 452–71.

Bartz, Jennifer A., et al. "Reminders of Social Connection Can Attenuate Anthropomorphism." Psychological Science, vol. 27, no. 12, 2016, pp. 1644–50, doi.org/10.1177/0956797616668510.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021, pp. 610–23, https://doi.org/10.1145/3442188.3445922 .

Brandtzaeg, Petter Bae, et al. "My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship." Human Communication Research, vol. 48, no. 3, 2022, pp. 404–29, doi.org/10.1093/hcr/hqac008.

Brown, Tom B., et al. "Language Models Are Few-Shot Learners." Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, arxiv.org/abs/2005.14165.

Carr, Nicholas. The Shallows: What the Internet Is Doing to Our Brains. W. W. Norton, 2010.

Christiano, Paul, et al. "Deep Reinforcement Learning from Human Preferences." arXiv, 12 Jun. 2017, arxiv.org/abs/1706.03741.

Deng, et al. "The Spillover of Incivility From Family to Workplace: The Mediating Effects of State of Self-Control." Atlantis Press, 2021.

Fang, Cathy Mengying, et al. "Investigating the Affective Use and Emotional Wellbeing on ChatGPT." MIT Media Lab, 21 Mar. 2025, www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/ .

Ferguson, Merideth. "You Cannot Leave It at the Office: Spillover and Crossover of Coworker Incivility." Journal of Organizational Behavior, vol. 33, no. 4, 2012, pp. 571–588.

Graybiel, Ann M. "Habits, Rituals, and the Evaluative Brain." Annual Review of Neuroscience, vol. 31, 2008, pp. 359–87.

"How People Are Using ChatGPT." OpenAI, 16 May 2026, https://openai.com/index/how-people-are-using-chatgpt/ .

Laestadius, Linnea, et al. "Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika." New Media & Society, vol. 26, no. 10, 2024, pp. 5923–41, doi.org/10.1177/14614448221142007.

Merrill, Jeffrey, et al. "Artificial Intelligence Chatbots as a Source of Virtual Social Support: Implications for Loneliness and Anxiety Management." Annals of the New York Academy of Sciences, 2025, doi.org/10.1111/nyas.15400.

OpenAI. “Aligning Language Models to Follow Instructions.” OpenAI, 27 Jan. 2022, https://openai.com/index/instruction-following/ .

OpenAI. “Introducing ChatGPT.” OpenAI, 30 Nov. 2022, https://openai.com/index/chatgpt/.

Pew Research Center. “Public Trust in Government: 1958–2024.” Pew Research Center, June 2024, www.pewresearch.org/politics/2024/06/24/americans-trust-in-federal-government-and-attitudes-toward-it/ .

Putnam, Robert D. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, 2000.

Reeves, Byron, and Clifford Nass. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, 1996.

Skjuve, Marita, et al. “My Chatbot Companion—A Study of Human-Chatbot Relationships.” International Journal of Human-Computer Studies, vol. 149, 2021, art. 102601.

Smith, Kyle S., and Ann M. Graybiel. “Investigating Habits: Strategies, Technologies and Models.” Frontiers in Behavioral Neuroscience, vol. 8, 2014, art. 39, doi.org/10.3389/fnbeh.2014.00039.

Turkle, Sherry. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, 2011.

Vaswani, Ashish, et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 5998–6008, arxiv.org/abs/1706.03762.

We Are Social, et al. “Number of Cumulative Monthly Active Users of ChatGPT Mobile App on Android and iOS from August 2023 to February 2026 (in Millions).” Statista, Statista Inc., 22 Apr. 2026, https://www.statista.com/statistics/1625099/monthly-active-users-of-chatgpt-mobile-app/

Weizenbaum, Joseph. Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman, 1976.

Designing the Ethical Being: Moral Enhancement, Autonomy, and the Limits of Irrevocable Consent By Jamie Lee

Abstract

Moral enhancement–the use of biomedical interventions to improve traits such as empathy, trust, and impulse control–raises a fundamental worry that intervening with our psychology will deprive us of our agency. These enhancements are not speculative: drugs such as oxytocin and selective serotonin reuptake inhibitors (SSRIs) are already used to influence emotional conditions under which moral decisions are made. Yet, this concern about agency is not straightforward because the threat to agency is not uniform across all forms of moral enhancement. Julian Savulescu and Ingmar Persson's defense of moral enhancement assumes that voluntary consent, in all cases, resolves any worries about agency. In this paper, I will argue that their assumption fails to account for significant differences between types of enhancement. I will do so by distinguishing between three levels of moral enhancement that threaten agency to varying degrees. These distinctions will reveal that consent alone cannot justify irrevocable enhancements that permanently remove the agent's capacity to reflect on and revise who they have become.

Introduction

Transhumanism is the scientific and philosophical movement premised on the claim that human nature has no fixed moral boundary. Nick Bostrom defines transhumanism as "the bold view that humans should exploit technological inventions that improve, lengthen, and yes, possibly change the lives of humankind" (Bostrom 3). On this view, human limitations are not sacred merely because they are natural. For Bostrom and many others, technology may permissibly improve upon limitations in lifespan, intellectual capacity, body functionality, sensory modalities, mood, energy, and self-control. One dimension of transhumanism, moral enhancement, extends this reasoning to moral psychology itself. Rather than enhancing only health, longevity, or intelligence, moral enhancements would aim to strengthen traits such as empathy, trust, and impulse control.

This paper concerns moral enhancement understood as interventions that aim to "increase the moral value of an agent's actions or character" (Agar 67). Moral enhancement must be distinguished from mere mood enhancement: a drug that makes a person happier does not necessarily make that person morally better.

There exists a crucial concern that moral enhancement will deprive us of our agency. This concern is often overgeneralized because different moral enhancements influence moral dispositions differently. To address this gap, I have structured this paper around three levels of commitment to moral enhancement, where "commitment" refers to the degree to which an agent gives up future control over their moral psychology. This structure avoids two mistakes. The first is the mistake of bioconservatism: the categorical rejection of moral enhancement simply because it modifies human nature (Browne 241). The second mistake is the consent fallacy: the

assumption that an agent's voluntary agreement to an enhancement renders it compatible with moral autonomy. Consent matters, but it is distinct from moral agency. Moral agency requires a metacognitive capacity: the ability to reflect on, evaluate, and revise one's motives and choices. Or, in other words, to think about one's thinking (Lai 4). The loss of this capacity would be a serious moral loss, even if the resulting behavior is improved.

What I am concerned about are enhancements that target our psychology through which moral evaluation occurs: not personal autonomy, the broader authority to shape one's own life, but moral autonomy, the capacity to reflect on moral reasons and to evaluate one's own actions (Christman). Moral enhancement threatens moral autonomy, which is dangerous not just because the agent loses options, but because the agent's moral compass is internally restructured in a way that prevents them from taking reflective distance.

In "Moral Enhancement, Freedom, and the God Machine," Julian Savulescu and Ingmar Persson offer a prominent argument for moral enhancement. They argue that ordinary traits such as impulsiveness, antipathy, short-term self-interest, and aggression are inadequate in such a technologically advanced world. Such moral failures, they argue, are serious enough to warrant biomedical interventions. If moral enhancement could reduce these failures, then intervention may not just be permissible. It may be necessary to prevent catastrophic outcomes (Savulescu and Persson 1).

However, Savulescu and Persson's argument has a gap. Although they worry that moral enhancement would deprive agents of their autonomy, they treat consent as a solution that works uniformly in all cases: they argue that consenting to any form of moral enhancement renders the enhancement an expression of agency rather than a violation of it (Savulescu and Persson 12). I argue instead that voluntary consent can justify some forms of moral enhancement, but cannot alone justify irrevocable enhancements that remove the conditions of future moral reflection. One cannot consent to the deprivation of moral agency.

Case 1: Moral Enhancements that Preserve Agency

The first level of moral enhancement comprises interventions that influence moral dispositions without fully removing the agent's future capacity for reflection and revision. These cases help demonstrate why moral enhancement should not be rejected categorically. Allen Buchanan's critique of worries about moral enhancement that appeal to human nature strengthens this point. Buchanan argues that a trait's naturalness does not show that it is good because human nature is a product of morally blind evolution shaped by reproductive fitness, not wisdom or rationality (Buchanan 16). If certain natural traits are harmful or obsolete, they should not be preserved on the sole grounds that they are natural. The relevant question is not whether an intervention modifies human nature, but whether a particular intervention is justified.

This point is not merely theoretical. Many morally valuable practices can alter an agent's values, emotional responses, and character. Education, for example, can change how we see right and wrong. Friendship can change what we value in others. Religious participation can reorient our priorities and sense of responsibility. If changing moral psychology alone were

objectionable, then we would have to condemn these practices we see as permissible, even admirable. Yet, none of these practices are thought to deprive agents of their moral autonomy.

Savulescu and Persson's examples of oxytocin and SSRIs fall within this first level of moral enhancement. They describe oxytocin as a drug that affects trust, cooperation, and generosity, and SSRIs as influencing aggression, fairness, and responses to harm (Savulescu and Persson 3-4). These examples illustrate why case 1 interventions are compatible with moral autonomy for two reasons. The first reason involves the nature of the intervention itself: oxytocin and SSRIs alter emotional conditions without restructuring the agent's capacity for moral reflection. The second feature is reversibility: the agent retains the ability to reflect on whether the intervention fits their values and whether to continue it. Together, the non-invasive nature and reversibility of case 1 enhancements preserve the agent's ongoing evaluative capacity. Therefore, if an agent identifies a specific trait as an obstacle to their well-being and moral commitments, then their voluntary adoption of reversible moral enhancement may support their agency rather than undermine it.

In cases that meet these conditions, I agree with Savulescu and Persson that consent provides sufficient justification to undergo moral enhancement. However, the difficulty arises when an enhancement is no longer reversible.

Case 2: Irreversible and Transformative Enhancements

The second level of moral enhancement involves interventions that are irreversible, transformative, or both. In these cases, the consent problem becomes more complicated, but not because irreversibility or transformation are themselves disqualifying for undergoing moral enhancement.

L.A. Paul argues that some experiences are transformative in one of two senses: they are either epistemically transformative, in that they provide knowledge one could not have acquired without the experience, or personally transformative, in that they redefine the agent's values, preferences, and identity. Parenthood is a standard example: one cannot fully understand what becoming a parent will be like before becoming one, and the experience may fundamentally change their priorities (Chan). Yet, such decisions can be reasonable exercises of autonomy because some decisions cannot be made solely based on rational calculation, since future preferences are unknowable in advance.

A similar point applies to transformative change more broadly. Consider a person convicted of a serious crime who undergoes intense justice rehabilitation. This process may permanently alter their understanding of harm, responsibility, and remorse in ways that they could not possibly have anticipated. Their moral outlook has been transformed. Yet this transformation does not undermine moral agency because the agent participates in the change through reflection, memory, and ongoing endorsement. The changed person is different but not deprived of their agency. Thus, transformation alone does not threaten moral autonomy.

Irreversibility alone is not the problem, either. Many irreversible decisions are central to living an autonomous life. Immigrating or choosing a vocation, for example, may close future

options, but do not close the capacity for reflection. A parent may be permanently changed, but the transformation typically unfolds through ongoing interpretation, endorsement, and revision. The parent can still ask if they endorse their new priorities, so irreversibility can preserve the metacognitive capacity necessary for moral agency.

A moral enhancement can therefore be irreversible while supporting autonomy. Consider a hypothetical neural intervention that strengthens the prefrontal cortex's inhibitory control over violent impulses. This enhancement does not implant new values or prevent the agent from reflecting on their actions; it simply strengthens the agent's capacity to act on reasons they endorse. The intervention is irreversible without being agency-replacing, and, in fact, can support the conditions of moral agency better than the agent's unmodified psychology. It is important to note that I am not endorsing people undergoing this procedure if it became available, since there are other important concerns. However, my argument is that the concern about moral agency does not apply to it despite our initial intuition that it does.

Case 3: Irrevocable, Agency-removing Enhancement

The third level of moral enhancement that I will argue is objectionable involves irrevocable, agency-removing interventions. Unlike cases 1 and 2, these enhancements do not merely alter dispositions. Instead, case 3 enhancements permanently restructure the agent's moral psychology and disable future reflection on their motivations and choices. We do not yet possess the technologies to perform these kinds of interventions, but that is precisely why these questions must be worked through now. If we develop the capacity for such enhancements before developing the philosophical frameworks needed to evaluate them, we will be dangerously unprepared.

Consider two hypothetical interventions. The first permanently eliminates aggressive desires. The second makes deception neurologically impossible, meaning the agent cannot form the intention to deceive. In both cases, the agent's behavior becomes less harmful, but this presumed improvement obscures a morally significant loss.

To recognize this loss, it is helpful to contrast these interventions with the previous two cases. In cases 1 and 2, the agent undergoes steady change, maintaining their capacity to evaluate the transformation and respond to their moral reflection. I previously agreed with Savulescu and Persson that consent can justify cases 1 and 2. However, the mistake they make is extending their consent-based justification for moral enhancement to this third kind of case as well.

The loss of this future reflection and the loss of the capacity to choose otherwise are related but distinct concepts. Moral reflection does not operate independently of the agent and their life decisions. Instead, it relies on the agent's capacity to feel a desire, to weigh it against alternatives, and to determine whether it is still worth acting on after this deliberation. When an intervention eliminates a desire, it forecloses these later steps of moral reflection. An agent who is incapable of forming violent intentions cannot reflect on the importance of restraint because they have lost the experience of choosing restraint over violence. An agent who cannot form the intention to deceive cannot fully understand the importance of honesty because they have not

chosen honesty over deceit. Therefore, this capacity for reflection depends on having live alternatives that enable agents to deliberate over. It is this optionality that preserves our autonomy.

Savulescu and Persson's God Machine falls into this third category. The God Machine is a hypothetical system from the year 2050 that "intervened in human action to prevent great harm, injustice or other deeply immoral behaviour from occurring" (Savulescu and Persson 10). This system would monitor the intentions of every human being and modify them within nanoseconds if the human were about to commit murder, rape, or other grossly immoral acts. Despite seeming like a breach of agency, "The God Machine was designed to give human beings near complete freedom", as humans would never consciously recognize the God Machine's interventions, and could still choose to be moral (Savulescu and Persson 10).

The God Machine's appeal is not difficult to understand: It preserves the liberty to make ordinary choices while preventing extreme wrongdoing. But at each decisive moment, the agent does not choose to act justly. The intervention instead bypasses deliberation, preventing them from choosing otherwise.

Here, the connection between agency-removing enhancements and the limits of consent must be articulated. The initial act of consenting to the God Machine is an exercise of moral autonomy: the agent has arrived at a judgement as a response to their reasons. But this consent gives moral authority to the God Machine, leaving not the same autonomous agent, but one without the capacity to reflect on whether a desire is worth acting on. The agent can no longer choose to live peacefully. Instead, they lose the capacity to choose otherwise, losing their future moral agency in the process.

Consent is not Moral Agency

The mistake that Savulescu and Persson make is treating consent as equivalent to autonomy. If a person voluntarily chooses an intervention, one may view that act as an expression of autonomy. Savulescu and Persson explicitly make this argument when they write, "There is one way in which the God Machine would not compromise autonomy, that is, even if it did prevent people from acting immorally. This would be the case if people voluntarily chose to be connected" (Savulescu and Persson 11). The thought is attractive: if autonomy means the power to make rational decisions, then a rational decision to prevent future wrongdoing may appear as an exercise of that very power.

But consent and moral agency are conceptually distinct. Consent is an act, while moral agency is a capacity. The act of consenting can express this capacity, but can also produce conditions that completely remove agency. Therefore, this paper's objection is not that consent is irrelevant for justifying moral enhancement. It is that, contrary to what Savulescu and Persson claim, consent by itself cannot justify interventions that threaten one's future capacity for reflection.

By moral agency, I do not mean perfect rationality and judgment. Human agents are inherently biased, confused, and mistaken at times. However, a minimum condition for moral

agency is the metacognitive capacity to take reflective distance from our motives and actions. We can determine if a desire is worth acting on, whether a past choice was wrong, or whether we act differently. Irrevocable, agency-replacing enhancements target this capacity, removing moral agency by removing moral deliberation.

The Ulysses Objection

The strongest objection to this paper's argument is that autonomy includes the power to bind oneself. After all, signing a contract, accepting a medical treatment, or entering a monastic order all impose long-term constraints on future choices. Yet, none of these acts are typically considered violations of autonomy; they are expressions of it.

Savulescu and Persson use Ulysses in an attempt to recombine consent and moral agency. In Homer's *The Odyssey,* Ulysses orders his sailors to bind him to the mast of his ship because he knows that when he hears the Sirens, he will be tempted to throw himself overboard (Savulescu and Persson 12). His self-binding protects him from future failure and is a rational use of autonomy. Savulescu and Persson invoke Ulysses to argue that consenting to a constraint on one's future agency is not only compatible with autonomy, but can itself be a judicious exercise of it. They write that "it is necessary to frustrate some of a person's desires, even his strongest desires, if we are to respect his autonomy" (Savulescu and Persson 12).

This objection is strong. If agents have authority over their future selves, then refusing irreversible moral enhancements might seem paternalistic. The objection only gets stronger when the enhancement would prevent harm to others, since the interests of a future victim may be more important than preserving the agent's autonomy. On this view, consenting to the God machine is no different in kind than Ulysses ordering his men to bind him with ropes. In both cases, the present self reasonably makes a precommitment on behalf of the future self to avoid bad judgment.

But this objection ultimately fails because the Ulysses case is not relevantly analogous to the God Machine. First, the ropes are external to Ulysses. They prevent him from acting on a passing temptation, but do not alter his internal moral psychology, values, or desires. When the ship passes the Sirens, the ropes are untied, and Ulysses is restored as a moral agent. He may reflect on his decision and endorse or revise his understanding of himself. The constraint acted only on his behavior, not him. The second reason is more fundamental than the first. Even if the ropes somehow did alter Ulysses internally, they only bind him temporarily. The restrictions end when the ship passes the Sirens, and whatever was changed could be reflected on afterwards. The God Machine is structurally different. It is not external, but instead restructures the psychology that generates behavior. It is also permanent: there is no point at which the intervention ends, leaving the agent restored. The God Machine threatens future agency itself, producing someone with better behavioral outputs, but without the capacity for genuine moral deliberation.

Conclusion

Savulescu and Persson are right to acknowledge that moral enhancement deserves serious attention. The gap between our technological powers and our moral psychology is real, and can be dangerous. Buchanan also raises an important point that evolution should not be viewed as a source of moral wisdom. Their arguments provide strong reasons to consider moral enhancement as worthy of philosophical attention rather than dismissible.

But the debate should not be framed as a choice between preserving human nature and embracing all enhancements. The real question is what it takes to justify a given moral enhancement. Some enhancements, such as SSRIs or oxytocin, do not threaten metacognitive capacities. These enhancements may influence moral dispositions without fully removing autonomy. Others are irreversible or transformative, but may still strengthen agency by allowing the agent to interpret and evaluate their new selves. These cases are more complex, but they resemble autonomous decisions that humans already make. Still, the most troubling cases permanently destroy the capacity for reflection. Consent alone cannot justify the permissibility of these moral enhancements.

The crucial objection to irrevocable moral enhancement is therefore not that it changes human nature. Nor is it that autonomy holds absolute power. Instead, the objection is that some forms of moral enhancement cannot be justified by consent alone. The goal of moral enhancement should not be limited to improving agent behavior or producing agents incapable of wrongdoing. It should be to help agents recognize moral reasons and act on their genuine values.

Ultimately, because of rapid technological progress, a safer world may be worth pursuing. But if this safety comes at the cost of losing moral agency, then we must reconsider our approach to moral enhancement.

Works Cited

Agar, Nicholas. "Why Is It Possible to Enhance Moral Status and Why Doing so Is Wrong?" Journal of Medical Ethics, vol. 39, no. 2, Aug. 2012, pp. 67–74, https://doi.org/10.1136/medethics-2012-100597.

Bostrom, Nick. "Transhumanist Values." Philpapers.Org, 2005, https://philpapers.org/rec/BOSTV. Accessed 9 July 2026.

Browne, Tamara Kayali, and Steve Clarke. "Bioconservatism, Bioenhancement and Backfiring." Journal of Moral Education, vol. 49, no. 2, Apr. 2019, pp. 241–56, https://doi.org/10.1080/03057240.2019.1576125.

Buchanan, Allen. Better Than Human. OUP USA, 2011.

Chan, Rebecca. "Transformative Experience." Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta and Uri Nodelman, Summer 2023, Metaphysics Research Lab, Stanford University, 2023, https://plato.stanford.edu/entries/transformative-experience/. Accessed 9 July 2026.

Christman, John. "Autonomy in Moral and Political Philosophy (Stanford Encyclopedia of Philosophy)." Stanford.Edu, 29 June 2020, https://plato.stanford.edu/entries/autonomy-moral/. Accessed 9 July 2026.

Lai, Emily. Metacognition: A Literature Review Research Report. 2011, https://nuovoeutile.it/wp-content/uploads/2015/11/Metacognition_Literature_Review_Final.pdf. Accessed 13 July 2026.

Savulescu, Julian, and Ingmar Persson. "Moral Enhancement, Freedom, and the God Machine." Monist, edited by Sherwood J. B. Sugden, vol. 95, no. 3, 2012, pp. 399–421, https://doi.org/10.5840/monist201295321.

Recovering the In-Between: Retrofitting Back Lanes and Pocket Gardens in Mong Kok, Hong Kong By Gunho Jacob Ahn

Abstract

Mong Kok is one of the most densely populated districts in the world; yet, it offers its residents almost no public green space. For the small in-between spaces it does have, the service back lanes and the official sitting-out areas are left neglected. This paper studies those two space types in the western part of the district through field observation, photography, and mapping on Google Maps and Street View. The fieldwork found that the back lanes work hard as storage, servicing, and shortcuts but fail as clean or safe places, while the pocket gardens keep greenery and rest apart, lack shade, and feel cut off from the street. In response, the paper develops two design proposals on a shared framework of blended functions and modular, vertical bamboo-scaffold construction. The first reorganises a back lane into a working ground floor with a planted promenade above, and the second revitalises a sitting-out area into a layered, communal garden. Both work with the activity already present rather than displacing it, directing the cooling and amenity benefits of green retrofits to the district's most heat-vulnerable users.

Keywords

Mong Kok, back lanes, pocket gardens, urban density, green space, modularity

Introduction

Step off one of Mong Kok's crowded shopping streets and into the gap behind it, and the city changes. The back lane is dark and narrow, lined with dripping air conditioners, stacked crates, and the smell of cigarettes, even though the flats above it are some of the most expensive per square foot on earth. This is the problem of Mong Kok. It is one of the most densely populated districts in the world, at around 130,000 people per square kilometre - a statistics that earned it a Guinness World Records listing as the busiest district in the world (*Overpopulation in Mong Kok*, 2023). Yet, the everyday spaces between its buildings, the back lanes and the small sitting-out areas, are neglected and underused.

This matters beyond one district. Hong Kong's planning standard asks for at least 2 square metres of open space per person. However, Mong Kok provides only 0.6 square metres per resident, among the lowest of any urban zone in the territory (Lai, 2017), and the wider Yau Ma Tei and Mong Kok area carries an acute 19-hectare shortfall of public open space against that standard (Urban Renewal Authority, 2022). Access to green space is a real shortfall in a city that keeps building upward. As dense cities grow, the pressure is usually to clear and replace, which removes the fine-grained urban structure and the informal life it supports. Thus, how can already dense cities recover value from the leftover spaces a dense district already has, rather than demanding land it does not have?

By examining the back lanes and the pocket gardens in Mong Kok, the study aims to identify current conditions and issues then generate designs that revitalize their original purpose

without displacing existing users. The paper moves from a literature review and site analysis to the field findings, then to a set of case studies, and finally to two proposals.

Preservation Versus Reconstruction: Jacobs and Le Corbusier

The core tension of modern urban planning theory is in the ideological conflict between preserving and retrofitting the existing built fabric versus demolishing it in favour of master-planned, highly structured modernist forms. This debate introduces fundamentally divergent methodologies for achieving urban vitality, structural organisation, and environmental resilience. Jane Jacobs's (1961) retrofit logic is utilised as this paper's primary orienting framework because Hong Kong's hyper-dense urban condition reflects her core concepts: its mixed-use fabric already structurally embodies many of her essential preconditions for an active urban ecology.

In *The Death and Life of Great American Cities*, Jacobs (1961) argues that intricate, dense, and socially diverse neighbourhoods create the foundation for a healthy city. To generate urban diversity, she establishes four indispensable structural conditions:

> 1. The district…must serve more than one primary function. 2. Most blocks must be short…streets and opportunities to turn corners must be frequent. 3. The district must mingle buildings that vary in age and condition. 4. There must be a sufficiently dense concentration of people. (Jacobs, 1961, pp. 150-151)

Crucially, Jacobs operates on a retrofit logic, claiming that urban issues should be solved by adapting existing infrastructure rather than completely rebuilding it, which helps maintain historical continuity, economic flexibility, and cultural identity. Her ideology connects to urban ecology, arguing that "only a genuine content of economic and social diversity…has meaning to the park and the power to confer the boon of life upon it" (Jacobs, 1961, p. 101). Open parks and green spaces cannot succeed as isolated pieces of land. Rather, their environmental and social connectedness depends entirely on the diversity of the surrounding urban fabric. Moreover, Jacobs critiques flawed, abstract statistical planning and the way it views communities, stating that "statistical people are a fiction for many reasons, one of which is that they are treated as if infinitely interchangeable" (Jacobs, 1961, p. 152).

Conversely, the modernist ideology pioneered by Le Corbusier (1929) rejected the "corridor-street" as he believed it was chaotic and harmful towards the environment. Additionally, he believed it limits natural light and traps dust and noise. He advocated for the absolute separation of urban functions, vertical density, and a "skyscraper-in-the-park." His theory is anchored by four fundamental principles: "1. We must de-congest the centres of our cities. 2. We must augment their density. 3. We must increase the means for getting about. 4. We must increase parks and open spaces" (Le Corbusier, 1929, p. 5).

Le Corbusier envisioned tearing down historical urban structures and replacing them with uniform, massive concrete towers and natural greenery. Under this view, urbanisation requires a

rejection of diverse, low-yield building typologies in favour of vertical architectural homogeneity. However, Le Corbusier converges with Jacobs on the critical necessity of introducing abundant green space and vegetation within the urban core. Their disagreement centres entirely on the means used to achieve this greenery: Le Corbusier believes that nature should be segregated into vast, abstract zones, whereas Jacobs envisions an integrate-and-retrofit logic that layers public parks and green roofs into the complex, pre-existing urban matrix.

Hong Kong's Inherited Density: The Composite Building

Hong Kong's inherited built environment presents a vivid example of this theoretical tension. The creation of this unique urban structure was heavily dictated by post-World War II socio-political realities. Facing massive, sudden population changes and a shortage of buildable flat land, Hong Kong's British colonial administration operated under a permissive, highly speculative, laissez-faire regulatory regime. This context is the driving factor of Hong Kong's modern extreme density (Seng, 2017).

Within this historical framework, the "composite building" - a hyper-dense, high-rise structure containing both domestic residences and diverse commercial or light-industrial operations - emerged as the city's dominant architectural typology. As shown in local urban histories, the composite building "exemplifies the paradox of collective sociability within an individual privatized space" (Seng, 2017, p. 82). Its rapid spread across Kowloon and Hong Kong Island was propelled by three primary drivers: the urgent requirement for spatially compact housing to shelter an escalating population, the rise of expanding middle-class investment pathways, and the explosive growth of a speculative domestic construction industry fuelled by integration of mainland Chinese migrants (Seng, 2017). Morphologically, the defining turning point occurred with the implementation of the revised 1955 Building (Planning) Regulations, which raised the height limit of buildings to "one and a half times the width of the adjacent road" (Seng, 2017, p. 87) while incorporating setback rules to protect street-level daylighting. This specific legal framework directly produced the famous silhouette of Hong Kong's skyline that characterises the city's urban infrastructure today.

The Urban Heat Island and Its Unequal Burden

Hong Kong's high density, canyon geometry, and heavily paved, impervious surfaces have collectively contributed to the rise of urban heat islands (UHIs). Wang et al. (2021) establish that the core mechanism of the UHI arises from "the modification of land surfaces with materials that effectively store short-wave radiation and reduce evaporative cooling" (p. 1), with anthropogenically generated waste heat from energy usage acting as a secondary contributor. Their results reveal that Hong Kong's UHI profile is "dominantly nocturnal during all months of the year, with the greatest magnitude in the winter and the smallest in spring" (Wang et al., 2021, p. 3). During the oppressive summer months, the nocturnal UHI intensity consistently hovers around 2-3°C in dense areas (Wang et al., 2021). Crucially for the theoretical argument for retrofitting, the leading predictors explaining local UHI responses are "impervious surface ratio,

surface albedo, street canyon aspect ratio, and vegetation density" (Kotharkar et al., as cited in Wang et al., 2021, p. 7). Because every single one of these variables is a distinct, physical attribute of the existing built fabric, they represent parameters that are fully actionable through tactical urban retrofits rather than full-scale demolition.

The consequences of this artificially altered microclimate are profound. As a result of its compact urban form and highly urbanised land cover, Hong Kong suffers from an intense UHI effect of up to 4°C (Tan et al., 2016). This severe heat level carries a quantifiable mortality link: each 1°C increase in temperature above a threshold of 28.2°C is associated with a 1.8% increase in mortality in Hong Kong (Chan et al., 2012). The burden, moreover, is unevenly shared. A 1°C rise above 29°C raises natural mortality by 4.1% in areas with a high UHI effect, compared with only 0.7% in low-UHI areas (Goggins et al., 2012), and under extreme heat the mortality risk of populations in high-UHI areas is roughly double that of those in moderate-UHI areas (Ho et al., 2023). This reality introduces a critical issue not previously acknowledged. Urban heat islands are not distributed evenly across Hong Kong. Environmental exposure and vulnerability are conditional on structural socio-economic variables, including localised housing typologies, subdivided building age, and a household's financial capacity to afford the continuous electricity costs required for mechanical air conditioning. Consequently, the core planning issue is no longer an analytical question of whether Hong Kong's inherited urban geometry generates hazardous heat, but rather a question of how this existing architectural form can be systematically retrofitted to protect its most vulnerable populations.

Green Infrastructure as a Retrofit Strategy

Recent research and Hong Kong government planning documents converge on a shared answer: the most viable path forward relies on integrating green and blue infrastructure directly onto the existing city, rather than completely replacing the infrastructure that already exists. This approach aligns closely with Jacobs's focus on organic preservation over modernist rebuilding.

This methodology is demonstrated in the government's *Hong Kong 2030+: Towards a Planning Vision and Strategy Transcending 2030*, specifically within its *Green and Blue Space Conceptual Framework* (Planning Department, 2016). The policy document acknowledges that Hong Kong is filled with green and blue assets, which must act as multi-functional systems to bolster urban resilience as the "lungs" of cities. The framework highlights that green infrastructure plays an "indispensable role in carbon sequestration, pollution amelioration, noise abatement, storm water management, and also relieving urban heat island effect" (Planning Department, 2016, p. 13). To counteract climate pressures and the rapidly ageing demographic profile, the policy updates land-use requirements by proposing an elevated planning target of 2.5 square metres of public open space per person (Planning Department, 2016). Furthermore, the framework advocates for a transition toward a "Biophilic City" model, mandating that urban design actively integrate natural elements into high-density pockets. This involves implementing soft waterfront edges to connect people to water bodies and using green infrastructure networks

to resolve climatic challenges without destroying the underlying cultural and historical identity of the neighbourhoods (Planning Department, 2016).

While these strategies lay out broad structural targets, contemporary green infrastructure and landscaping literature positions the streetscape as the primary, most effective lever for executing these retrofits. Streetscapes are prioritised because they function simultaneously as critical mobility conduits and major heat stores (Fallast et al., 2021). Introducing vegetation within these corridors represents a powerful method for cooling street-level temperatures and mitigating flash floods. This tactical approach is increasingly advanced through the concept of green acupuncture, which implements small and highly localised urban interventions, such as pocket parklets, green walls, or mobile urban tree containers, that insert natural elements directly into dense urban cracks without requiring mass demolition (Fallast et al., 2021). The efficacy of these interventions is supported by research demonstrating that just one square metre of extensive green roof is capable of evaporating up to two litres of water daily, absorbing 375 grams of carbon dioxide, and binding ten grams of airborne fine dust annually (Fallast et al., 2021). Within the Hong Kong context, these modular green roof metrics are directly transferable to the vast, underutilised flat rooftops of the traditional composite buildings analysed previously, offering an immediate path toward modification of Hong Kong's climate.

Additionally, local data provide the quantitative boundaries required to maximise the cooling performance of these green retrofits within Hong Kong's specific microclimate. Parametric modelling indicates that tree planting covering roughly one third of the urban surface is needed to lower pedestrian-level air temperature by around 1°C (Ng et al., 2012), and a Hong Kong government consultancy guideline recommends that 20% to 30% of building site areas be provided with greenery (Ng et al., 2012). Crucially, this research demonstrates that spatial placement is just as critical as the absolute volume of vegetation. The cooling of air temperature and sensible heat is twice as high for vegetation arranged in active urban wind corridors as for identical vegetation volumes placed in stagnant, sheltered leeward pockets, and the local sky view factor (SVF) serves as the primary morphological predictor for determining where urban trees will deliver the greatest temperature reduction (Tan et al., 2016). Together, Hong Kong's municipal policy frameworks, the streetscape retrofit literature, and specific microclimate research establish a coherent retrofit-based strategy. However, while this technical framework successfully connects urban form with environmental mechanics, it fails to address the underlying structures of social equity.

The Equity Gap in Existing Research

While the current body of literature firmly establishes that retrofit-based green infrastructure can measurably mitigate Hong Kong's urban heat island effect, and conclusively demonstrates that Hong Kong's historic, high-density urban structure is the underlying cause of the urban heat islands, a critical gap remains. The existing research has failed to structurally connect these findings to the deeply unequal distribution of heat exposure across Hong Kong's diverse housing types and income groups. The essential equity dimension of climate adaptation

remains largely absent from mainstream urban design studies. Current research treats urban populations as uniform bodies, the "statistical people" that Jane Jacobs criticised as a planning fiction. As a result, the literature has neglected to investigate which specific demographics will actually benefit from newly retrofitted urban green infrastructure, and which marginalised communities will continue to bear the disproportionate physical burden of localised heat storage.

This paper directly addresses this literature gap by pairing the technical retrofit literature with ground-level fieldwork in a district where heat exposure and open-space deprivation converge most sharply. It documents who actually uses the city's leftover spaces, the back lanes and sitting-out areas of Mong Kok, and proposes designs that extend the benefits of green retrofits to those existing users, including elderly residents, informal workers, and cardboard collectors, rather than displacing them.

Site Selection: Why Mong Kok

Mong Kok is a small district in Kowloon and one of the most densely built places anywhere (see Figures 1 and 2). The wider Yau Tsim Mong district that contains it held 310,647 residents across 6.99 square kilometres in the 2021 census, about 44,500 people per square kilometre, so Mong Kok's own core sits far above even that average (Census and Statistics Department, 2021). That density comes from how the district stacks uses on top of each other, with wet markets and street stalls at street level, shops and restaurants on the ground floor, and flats in the same buildings above. The street pattern is still the old fine grid, with main commercial streets and a web of narrow back lanes threaded behind them.

This is what makes Mong Kok a good place to study how tight urban space can be reworked. The density means that every leftover slot, whether a back lane or a gap between buildings, gets used for something, formally or not. Green space is the clearest shortfall. Hong Kong's own planning standard, the Hong Kong Planning Standards and Guidelines, asks for at least 2 square metres of open space per person, and Mong Kok sits far below that at 0.6 square metres per person, the worst provision of any urban zone in the territory (Lai, 2017). So the question is not whether to build new parks on empty land, because there is none, but how to recover the small, overlooked spaces the district already has.

Figure 1

Location of Mong Kok Within Hong Kong

Figure 2

Site Map of the Mong Kok Study Area

The study boundary sits inside 0.25 square kilometres, where the density is 130,000 people per square kilometre. Within it, eleven public green or sitting-out areas were identified, a small number for an area this dense. Figure 3 maps the surveyed back lanes and sitting-out areas by condition.

Figure 3

Site Condition Map of Surveyed Back Lanes and Sitting-Out Areas

Methodology

Evidence was gathered through field observation, photography, and research on Google Maps and Street View.

Field Visits

The visit took place on June 4, in the afternoon between approximately 1 p.m. and 4 p.m. The fieldwork focused on the western side of Mong Kok, the cluster of streets around

Reclamation Street, Canton Road, Soy Street, Nelson Street, Shanghai Street, Shantung Street, and Yin Chong Street, between the Mong Kok and Yau Ma Tei MTR stations. This area was chosen because it holds both conditions close together, with working back lanes behind the shops and an array of sitting-out areas. As a result, they could be compared on foot in a single loop.

Observation and Selection

At each site, photographs were taken and notes recorded on the physical condition, who was using the space, and roughly when. For the back lanes, the notes covered the geometry, the storage and servicing, the air-conditioner placement, drainage, smell, and lighting. For the gardens, they covered seating, shade, greenery, paving, lighting, and how cut off the space felt from the street. Not every lane and garden in the district was studied, but five unique lane geometries were identified (parallel, intersection, curve, opening to open space, and cul-de-sac), along with three garden states (poor, good, and an empty cul-de-sac that could become one).

Mapping

Google Maps was used to set the boundary and locate sites, and Street View to read context. Because its imagery is dated, Street View also allows a comparison of past and present conditions, and the same imagery underpins the diagrams and visual material presented in this paper.

Limitations

The method has real limits. The fieldwork was carried out alone over only a few hours, so it caught only certain days and times and missed seasonal or late-night patterns. No structured interviews were run, so what is said about who uses these spaces, and why, comes from direct observation rather than from talking to users. Street View only shows what the camera car could reach, which leaves out the interiors of the tightest lanes and cul-de-sacs.

Field Research Findings

The fieldwork focused on Mong Kok and on two everyday space types the district depends on but rarely designs for: the back lanes that run between the buildings and the sitting-out areas tucked into the block (see Figure 3). The findings are kept to these two conditions, and because the gardens vary so much, they are split into the ones left to decay and the ones that work. The sections below set out what each space looks like, who uses it, when, and where it helps or fails the people nearby. Table 1 summarises the eight sites documented in detail.

Table 1

Summary of Sites Documented During Fieldwork

Site	Type	Condition	Key features observed	Main users observed
Cheong Sha Street back lane	Back lane	Working, degraded	Storage, servicing, air-conditioner condensers; no greenery; poor drainage; unlit at night	Shopkeepers, kitchen staff, delivery workers, cardboard collectors
Shanghai St/Shantung St	Sitting-out area	Poor	One mature tree, fixed benches, hard paving, dim lighting; cut off from street	Older men resting, smokers, delivery riders
Reclamation St	Sitting-out area	Poor	No shade, minimal planting, lifeless	Occasional midday resters
Shan Tung St	Sitting-out area	Poor	Ornamental planting only; no proper seating	Few users
Canton Rd/Soy St	Sitting-out area	Poor	Mid-block, set back from pavement; little reason to enter	Few users
Reclamation St/Soy St	Sitting-out area	Poor	Mid-block, set back; fixed seating	Few users
Reclamation St/Nelson St	Sitting-out area	Good	Varied seating, exercise machines, playground, ramps, lighting; doubles as walkway	All ages, day and evening
Canton Rd/Nelson St	Sitting-out area	Good	Similar features; on walking routes, open to the street	All ages

Note. Attributes reflect direct observation during a single field visit (June 4, approximately 1-4 p.m.).

Condition 1: Back Lanes

The back lane is a service space built into the old street grid so crews could haul waste out from behind the buildings, doubling as a fire corridor. In Mong Kok these lanes sit behind the shopfronts, running parallel to the commercial streets and reachable through narrow gaps off the pavement. They are tight, barely wide enough for a person and a handcart to pass, and run the full depth of a block. Inside are gas pipes, dripping air conditioners, cardboard towers, restaurant crates, and graffiti (see Figure 4).

Figure 4

The Cheong Sha Street Back Lane

The lanes are sorted into five conditions by their geometry: straight lanes, intersections, curved lanes, lanes that open out into a wider space, and dead-end cul-de-sacs. The one selected for redesign is a parallel back lane off Cheong Sha Street, near Hotel MK and a block from the Shanghai Street/Shantung Street sitting-out area, which pairs the alley with a nearby garden. Use shifts by the hour. Shopkeepers store stock here, kitchen staff prep and step out for smoke breaks, delivery workers cut through, and elderly cardboard collectors fill their carts. The busiest moments are the early-morning deliveries and the meal-service rush.

A few things genuinely work. The lanes give cramped shops and kitchens the storage their frontages cannot hold. They are a real shortcut for people on foot and for deliveries. They carry small social moments for workers stepping out of the heat. They also hide the air-conditioner condensers off the main facade and keep that heat away from the busy streets, and the open slot brings light and air to the flats above and acts as a light well.

The same lanes fail as anything clean or safe. Sanitation is the first thing a visitor notices, the smell and the trash, and it rarely gets cleaned up. The tight alleys trap Hong Kong's humidity and heat, made worse by water dripping from the units overhead. There is no greenery. Drainage is poor, so the floor floods, and at night there is almost no lighting, so most people would not walk in.

Condition 2: Pocket Gardens in Poor Condition

Jacobs (1961) argued that a park is not good on its own; it works only when the streets and uses around it give people a reason to be there. The weaker sitting-out areas visited make her point for her. These are the city's official rest gardens, wedged between buildings or under flyovers, with benches, a few planters, paving, a bin, and sometimes a mural or a water-refill station. What is usually missing is shade and any real greenery, and in practice they serve more as smoking spots than shared space (see Figures 5-9).

The Shanghai Street/Shantung Street Sitting-out Area is the one selected for redesign, and it shows the problem clearly. It has a single large tree, a line of fixed benches, and hard paving with dim lighting. There is little usable shade away from the tree, the seating cannot be moved, and the space feels cut off from the street rather than part of it. The Reclamation Street Sitting-out Area is arguably as bad, with its lifeless character and lack of any shading. The Shan Tung Street Sitting-out Area has planting that can only be looked at, with nowhere proper to sit, and the gardens at Canton Road/Soy Street and Reclamation Street/Soy Street are much the same. They sit mid-block, set back from the pavement, so pedestrians have no reason to go in. The people who use them are mostly older men resting alone, smokers, and delivery riders waiting on an order, and they come around midday and late afternoon for a quick sit. By evening the spaces are dead, because they feel exposed and a little unsafe once the shops close. The seating is fixed, the floor is asphalt or concrete, and the garden turns its back on everything around it.

Figure 5

Shanghai Street/Shantung Street Sitting-Out Area

Figure 6

Reclamation Street Sitting-Out Area

Figure 7

Shan Tung Street Sitting-Out Area

Figure 8

Canton Road/Soy Street Sitting-Out Area

Figure 9

Reclamation Street/Soy Street Sitting-Out Area

Condition 3: Pocket Gardens in Good Condition

The stronger gardens follow Jacobs more closely, because they are busier as more kinds of people have a reason to be there. The Reclamation Street/Nelson Street Sitting-out Area is the clearest good example, with the garden at Canton Road/Nelson Street close behind (see Figure 10). They are larger and more varied than the poor ones, they sit along routes people already walk, and they open onto the street rather than hiding from it. They work because they give people something to do at any age. There is plenty of seating, comfortable benches, simple exercise machines, and a small playground for children. Ramps make them reachable for everyone, and proper lighting keeps them in use after dark. The planting runs through the space instead of being fenced off to one side, and the garden often doubles as a walkway, so it stays active. They still fall short. Shade is limited. Even where greenery is present, it is often there to

look at rather than to use, and it eats into space that could be seated. The benches are mostly fixed again, so people cannot angle or move them. Maintenance and hygiene are still an issue, especially after rain, when parts of the ground flood and stay wet. And for all the activity, these gardens still do little to connect with the buildings and uses right beside them.

Figure 10

Reclamation Street/Nelson Street Sitting-Out Area

Case Studies

Across the two conditions there were many noticeable flaws: back lanes that work hard but feel hostile, pocket gardens that are static and cut off, and almost no spare ground to build on. As a result, the search turned to projects that find ways to make a service lane liveable while it keeps working, ways to go vertical and modular given the density, and ways to mesh greenery and walkways. The five references below are grouped by the concept drawn from each.

Concept 1: Alleyway Activation

Jade Alley, by Daniel Toole Architecture and SB Architects in Miami, turns a leftover service alley into a public room, lined with repeating arched arcades that give the narrow slot a clear identity (*Jade Alley / Daniel Toole Architecture*, 2024). *G(U)ARDEN*, by Annvil in Riga, is a 54-square-metre experiment that hangs a vertical garden of crate planters on a light scaffold frame against a blank wall (*G(U)ARDEN Vertical Urban Garden / Annvil*, 2020).

Both speak directly to the Cheong Sha Street back lane. Jade Alley shows how a repeating overhead frame can turn an empty slot into a place without widening it, which matters where the lane is barely 2.5 to 3 metres wide. G(U)ARDEN shows how to green a lane using only scaffolding, since there is no floor to give up. What cannot be copied is Jade Alley's clean and empty starting point. The Mong Kok lane has to stay a working service route, so any structure must be accessible, and certain elements must be demountable, with areas kept for storage.

Figure 11

Jade Alley, Miami

Note. From *Jade Alley / Daniel Toole Architecture*, 2024, ArchDaily (https://www.archdaily.com/1018908/jade-alley-daniel-toole-architecture-plus-sb-architects).

Figure 12

G(U)ARDEN Vertical Urban Garden, Riga

Note. From *G(U)ARDEN Vertical Urban Garden / Annvil*, 2020, ArchDaily (https://www.archdaily.com/948897/g-u-arden-vertical-urban-garden-annvil).

Concept 2: Vertical and Modular Gardens

Dyv-net, a Hong Kong proposal by JAPA Architects, stacks farming and public space into a slim vertical tower, using a light metal structure and shifting floor plates modelled on terraced rice farming (Furuto, 2013).

The poor sitting-out areas, including the Shanghai Street/Shantung Street garden selected for redesign, fail because everything sits on one flat, shaded level. Dyv-net's lesson is to use the vertical. Stacking lets a small area hold both movement and rest. A ramp or stair that climbs becomes the route, and the flat ground becomes a place to sit. The terraced plates carry planting at every level, so green and seating share space rather than compete for it. It also argues for keeping a working layer, productive growing, so the pocket garden does something beyond a quick sit.

Figure 13

Dyv-net, Dynamic Vertical Networks Proposal, Hong Kong

Note. From *Dyv-net, Dynamic Vertical Networks Proposal / JAPA Architects*, by A. Furuto, 2013, ArchDaily (https://www.archdaily.com/401343/dyv-net-dynamic-vertical-networks-proposal-japa-architect).

Concept 3: Walkway and Garden Integration

The Next Hutong, by MVRDV in Beijing, regenerates dense alley fabric plot by plot, threading pocket gardens through the lanes as a kit of small, repeatable interventions rather than one master plan (Rosenfield, 2015). *Hop Park*, by Shma in Bangkok, is a modular green deck that pops up over existing infrastructure with the least possible change to what is already there, with unique seating elements that enhance the park's identity (*Hop Park, Green Pop-up Hop into the Greeniverse / Shma Company Limited*, 2024). The good pocket gardens already mix park and walkway, and these two show how to push that further here. Both work on narrow, awkward sites by going light and modular, which suits a city that already raises tall temporary bamboo scaffolds and has an abundance of wasted space. Planting and seating can be more integrated and movable, freeing the ground for movement. Modularity matters most for maintenance, the weak point observed after rain, because demountable, swappable parts can be cleaned or replaced one unit at a time.

Figure 14

The Next Hutong, Beijing

Note. From *MVRDV Reimagines the Chinese Hutong*, by K. Rosenfield, 2015, ArchDaily (https://www.archdaily.com/779706/the-next-hutong-mvrdv).

Figure 15

Hop Park, Bangkok

Note. From *Hop Park, Green Pop-up Hop into the Greeniverse / Shma Company Limited*, 2024, ArchDaily (https://www.archdaily.com/1020099/hop-park-green-pop-up-hop-into-the-greeniverse-shma-company-limited).

Design Proposals

The fieldwork produced two design proposals, one for the back lane and one for the pocket garden (see Figure 16). Together they answer all three site conditions, since the garden proposal draws its lessons from both the poor gardens being fixed and the good ones that already work. Both run on the same framework. First, blend functions instead of separating them, so a working lane can also be green and restful, and a rest garden can also hold people of all ages and nature. Second, build modularity into both conditions, with demountable parts on vertical bamboo scaffolding, so the design fits a site with no spare ground and can be cleaned or swapped one piece at a time. Each proposal takes on the failures documented in the field research findings and leans on the case studies discussed above.

Figure 16

Detailed Site Map of the Two Proposal Sites

Proposal 1: Back Lane Activation

The site is the Cheong Sha Street back lane, barely 2.5 to 3 metres wide. The lane keeps its working life at ground level, since storage, deliveries, air-conditioner units, and kitchen prep are the reason it exists. The first move is to organise that work. The floor storage becomes a wooden fenced enclosure that runs along the service side, taking up only half of the width of the alleyway, with a door for the shops to load and unload, which lifts the stock off the floor that floods after rain and organises the messy area. The other half becomes a quick passage through the alleyway, making deliveries more efficient because the pathway is free of trash, and giving pedestrians a shortcut to their destination that is not overflowing with rubbish and puddles of water. Figure 17 diagrams the organising concept, and Figure 18 shows the ground level.

Figure 17

Concept Diagram of the Back-Lane Activation

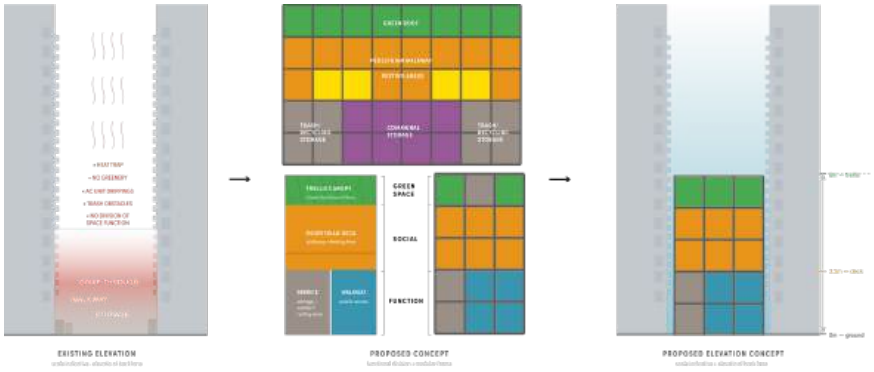


Figure 18

Elevation Rendering of the Proposed Alleyway Revitalisation, Ground Level

Another pedestrian route rises above the ground level. A raised deck spans the full width of the lane, wall to wall, so that it gives shading for the storage and pedestrian walkway at the ground floor. The second level runs the lane's length as an elevated promenade set under a planted bamboo canopy. This gives a covered footpath the full 3.3 metres wide, with the greenery high overhead and green vines on the sides of the buildings, so the walking clearance

sits well past the 2.0-to-2.3-metre minimum. Light benches line both walls and leave a clear central path far wider than the 1.1-metre minimum. The deck is reached by a stair at each end, one rising and one falling, so it reads as a through-route rather than a dead end. Each flight is eleven steps, under the sixteen-step limit before a landing is required, and keeps the 2.0-metre clear height overhead. The staircases are the width of the storage area so they do not block the smaller pedestrian pathway underneath (see Figure 19).

Figure 19

Elevation Rendering of the Proposed Alleyway Revitalisation, Elevated Promenade

The people who already use the lane keep their ground, the shopkeepers, kitchen staff, delivery workers, and the elderly gathering cardboard, while residents and passers-by gain a green, shaded promenade above the working floor. By day the ground stays busy with deliveries and meal service while the deck carries a quiet, planted walk. At night, lights along the scaffold keep the promenade open and overlooked. This follows Jacobs's (1961) idea of eyes on the street: keeping the lane used and watched through the day is what makes it feel safe after dark, rather than fencing it off. Jade Alley shows that a repeating overhead frame can give a narrow slot identity without widening it, and G(U)ARDEN shows that the blank walls, not the floor, are where the greenery has to go.

Proposal 2: Pocket Garden Revitalisation

The site is the Shanghai Street/Shantung Street Sitting-out Area. Its core failure is that greenery and seating are kept apart, there is no real shade, the benches are fixed, and the space feels rigid rather than communal. The mature central tree stays and becomes the centre of the design, now growing from a sunken rain-garden basin that takes the water flooding the paving and turns it into planting. Two curved ramps mirror each other around the tree and climb from the ground to a raised rear terrace, replacing the old steps so the garden is reachable for everyone and the climb itself becomes the route. Figure 20 diagrams the organising concept. From the terrace, a tight spiral stair winds up the trunk to a small deck set among the canopy, so the garden uses its full height instead of one flat level (see Figures 21 and 22).

Figure 20

Concept Diagram of the Proposed Pocket Garden Revitalisation

Figure 21

Elevation Rendering of the Proposed Pocket Garden Revitalisation, Version 1

Figure 22

Elevation Rendering of the Proposed Pocket Garden Revitalisation, Version 2

Onto this frame plugs a modular kit. Green-roofed pillar canopies and light umbrellas give shade, vertical columns carry plants up the blank firewalls and double as lighting, and the seating is loose so people can move it where they like. The ground and terrace are programmed for more than a quick sit, with a flexible platform for morning tai chi or Sunday gatherings, a fitness corner, a small playground, and communal garden plots that give the working layer a purpose. A fence and a planted entrance arch open the garden to the street instead of hiding it. In the day it holds elderly residents resting, people exercising, children playing, and neighbours tending plots; in the afternoon it takes gatherings; at night the lit columns keep it open and feeling safe. This is Jacobs's (1961) point about parks, that they only work when the uses around and within them give people a reason to come. Dyv-net shows how stacking turns a tiny footprint into both movement and rest, and Hop Park and The Next Hutong show how a modular, demountable kit keeps the space active and maintainable where maintenance is the weak point observed after rain.

Conclusion

Back lanes and pocket gardens are underused assets in Hong Kong's high-density, low-green-space condition, and the better response is to work with the activity already in them rather than to clear and replace it. This paper has shown that Mong Kok's back lanes already carry storage, servicing, social breaks, and shortcuts, and that its sitting-out areas already hold the basic idea of public rest, so the problem is not absence but neglect and rigidity. Jacobs's (1961) argument that a public space works only when the uses around and within it give people a reason to be there was visible across the fieldwork, in the busy lanes that still felt unsafe at night and in the empty gardens cut off from the street. The proposals also make the paper's theoretical synthesis explicit: rather than choosing between the two thinkers, the back-lane scheme borrows Le Corbusier's (1929) vertical separation of pedestrians from servicing while keeping Jacobs's street-level life intact below it, layering the modernist section onto the preserved fine grain instead of replacing that grain.

The two design proposals follow from this. The back-lane scheme keeps the working ground floor and lifts a planted pedestrian promenade above it, while the pocket-garden scheme weaves greenery, movement, and rest together on a layered, accessible structure. Both run on the same principle of blending functions and building from a modular, vertical bamboo-scaffold kit that suits a site with no spare ground. In answer to the research question, the proposals suggest that the district's existing infrastructure of lanes and pocket gardens can host enough added greenery to contribute meaningfully to street-level cooling, and that, because their existing users are among the district's most heat-vulnerable residents and workers, improving these spaces in place directs that cooling to the people who need it first.

The work has clear limits. It rests on 3D models rather than detailed or tested construction, its cooling estimates are indicative rather than simulated, and it was carried out without community engagement. The next step is to test these ideas with the people who use these spaces, and to grow a Mong Kok where its forgotten in-between spaces become part of its everyday green network.

Works Cited

Census and Statistics Department. (2021). 2021 population census: Yau Tsim Mong district statistics [Data set]. Centamap. https://census.centamap.com/hong-kong/Yau%20Tsim%20Mong

Chan, E. Y. Y., Goggins, W. B., Kim, J. J., & Griffiths, S. M. (2012). A study of intracity variation of temperature-related mortality and socioeconomic status among the Chinese population in Hong Kong. Journal of Epidemiology and Community Health, 66(4), 322-327. https://doi.org/10.1136/jech.2008.085167

Fallast, M. T., Pansinger, S., Krebs, G., Moser, M., & Zobl, A. (2021). Systematically retrofitting city streets: Meeting the demands of climate change through multifunctional climate-responsive street gardens. Urbani izziv, 32(1), 111-122. https://doi.org/10.5379/urbani-izziv-en-2021-32-01-004

Furuto, A. (2013, July 13). Dyv-net, dynamic vertical networks proposal / JAPA Architects. ArchDaily. https://www.archdaily.com/401343/dyv-net-dynamic-vertical-networks-proposal-japa-architects

G(U)ARDEN vertical urban garden / Annvil. (2020, October 6). ArchDaily. https://www.archdaily.com/948897/g-u-arden-vertical-urban-garden-annvil

Hop Park, green pop-up hop into the greeniverse / Shma Company Limited. (2024, August 15). ArchDaily. https://www.archdaily.com/1020099/hop-park-green-pop-up-hop-into-the-greeniverse-shma-company-limited

Jacobs, J. (1961). The death and life of great American cities. Vintage Books.

Jade Alley / Daniel Toole Architecture. (2024, July 16). ArchDaily. https://www.archdaily.com/1018908/jade-alley-daniel-toole-architecture-plus-sb-architects

Lai, C. (2017). Unopened space: Mapping equitable availability of open space in Hong Kong. Civic Exchange. https://civic-exchange.org/wp-content/uploads/2017/04/20170224POSreport_FINAL.pdf

Le Corbusier. (1929). The city of to-morrow and its planning (F. Etchells, Trans.). Payson & Clarke.

Ng, E., Chen, L., Wang, Y., & Yuan, C. (2012). A study on the cooling effects of greening in a high-density city: An experience from Hong Kong. Building and Environment, 47, 256–271. https://doi.org/10.1016/j.buildenv.2011.07.014

Overpopulation in Mong Kok. (2023, May 22). ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/685eb01e1328412d9b98a7afbde0b4b0

Planning Department. (2016). Hong Kong 2030+: Towards a planning vision and strategy transcending 2030 — Green and blue space conceptual framework. Hong Kong SAR Government.

https://www.pland.gov.hk/file/planning_studies/comp_s/hk2030plus/document/Green%20and%20Blue%20Space%20Conceptual%20Framework_Eng.pdf
Rosenfield, K. (2015, December 31). MVRDV reimagines the Chinese hutong. ArchDaily. https://www.archdaily.com/779706/the-next-hutong-mvrdv
Seng, E. (2017). The city in a building: A brief social history of urban Hong Kong. Studies in History and Theory of Architecture, 5, 81-98. https://doi.org/10.54508/sITA.5.07
Tan, Z., Lau, K. K.-L., & Ng, E. (2016). Urban tree design approaches for mitigating daytime urban heat island effects in a high-density urban environment. Energy and Buildings, 114, 265-274. https://doi.org/10.1016/j.enbuild.2015.06.031
Urban Renewal Authority. (2022). Summary of findings and recommendations of the District Study for Yau Ma Tei and Mong Kok (TPB Paper No. 10795). Town Planning Board. https://www.tpb.gov.hk/uploads/page/meetings/TPB/Yau_Mong_Study/Yau_Mong_Study_MainPaper.pdf
Wang, W., Chen, B., Xu, Y., Zhou, W., & Wang, X. (2021). Urban heat islands in Hong Kong: Bonding with atmospheric stability. Atmospheric Science Letters, 22(6), Article e1032. https://doi.org/10.1002/asl.1032

Improving Snake Robots Terrain Capability with a Passive Gripping Mechanism
By Naman Gujral

Abstract

Snake robots have become increasingly valuable in applications such as search-and-rescue, the medical field, and exploration because of their ability to navigate tight and uneven environments. However, existing snake robot designs continue to face challenges in steep and rough terrain due to traction limitations and reduced stability. There are many current approaches to improve terrain adaptability, such as anisotropic frictional scales, which improve directional grip but do not increase normal force, and modular locomotion, which uses control algorithms rather than mechanical traction. Therefore, a passive adaptive gripping mechanism is proposed to combine rubberized anisotropic frictional scales and deployable flexible brushes to increase ground contact, reduce slippage, and maintain stability while limiting additional energy consumption. This combines multiple passive strategies and offers a theoretical solution for snake robots traveling over unpredictable terrain, as well as a foundation for future prototypes.

Introduction

Since the first snake robot was developed in 1971 by Shigeo Hirose (Hirose and Yamada), the field has seen significant advancements in both locomotion and reliability, particularly in replicating the natural movement of real snakes. Today, snake robots are being experimented and applied to many different fields such as the medical for surgical applications (Wang et al.) and in search and rescue missions (Han et al.). Snake robots have flexible, multiple joint bodies that allow them to move through uneven environments unlike traditional robots (Hirose and Yamada). Due to these advantages, researchers are continuing to explore new ways to make snake robots more effective for real-world use.

In recent advancements, snake robots have focused on a variety of new actuation systems that help their locomotion better replicate how real snakes move. For example, tendon-driven systems use advanced cable actuation that reduces the number of active segments while still allowing the robot to generate smooth, wave-like motion, like a real life snake (Transeth et al.). Additionally, Vacuum-powered actuators use modular systems inspired by snakeskin to control friction depending on the direction of movement, which helps the robot grip specific surfaces more effectively (Lee et al.). These actuation systems allow snake robots to adjust their movement based on the terrain they are on.

Along with development of snake robot movement, there have been other improvements in materials and structural design. For instance, many robots use soft or flexible materi-als, which improve terrain adaptability and make them safer around people (Transeth et al.). Additionally, there are other mechanical improvements including the use of modular parts that can bend in multiple directions to better mimic the movement of a real snake and driving assistant mechanisms that help prevent the robot from rolling while increasing movement speed (Lee et al.). Overall, all these innovations are helping snake robots reach higher efficiency of

locomotion, adaptability, and performance in real-world environments.

Even though snake robots have become more advanced, they still face problems when it comes to traveling over rough or steep terrain. Current designs struggle with getting enough grip and torque to stay stable while climbing or moving across loose or uneven surfaces. This is a significant issue as snake robots need to move seemingless through rubble, rocks, or narrow vertical spaces (Liu et al.). Without the right amount of traction, snake robots can slip, or use constant power than needed, resulting in an ineffective locomotion. This paper will explain how an adaptive gripping mechanism can improve a snake robot's climbing ability and stability on rough or elevated terrain by overcoming grip and torque limitations.

Limitations with Previous Designs

Snake robots have demonstrated efficiency in controlled environments, but their performance varies significantly when moving across rough or inclined terrain. For example, one modular soft snake robot experienced a 33.

Researchers have explored several solutions to improve snake robots' ability to handle uneven terrain. One approach is a snakeskin-inspired 3D printable soft robot, introduced with anisotropic friction scales that can create higher resistance in the backward direction. This improves the traction problem on slippery surfaces and allows accurate locomotion on glass up to 30 degrees (Lee et al.). Although the anisotropic scales increase directional grip, they don't passively amplify normal force or redistribute the surface contacts for increased traction, so the snake robot still requires high torques to climb on irregular terrain, while the slippage issue may be mitigated.

A second approach is a 12 degree of freedom modular snake robot with alternating joint orientations; this addresses stability and adaptability by introducing bio-inspired gait options. One movement option is Sidewinding, which achieves a reduced heading error of -1.26 with a 2.88 degree variation, and smoother lateral movement on uneven surfaces. This design addressed the directional instability found in earlier robots, and demonstrates efficiency in climbing through careful gait selection to adapt to different terrains (Marin and Vivas). Although the selected movement provides an accurate locomotion for real-life snake climbing, this solution may rely solely on control strategies and may be limited when an efficient grip and torque are insufficient.

The third approach is a body compliance, focused on obstacle traversal by enabling the robot's body to deform out of plane and move over large obstacles. This enhanced stability and allowed the robot to distribute the surface forces more evenly (Fu and Li). While this mechanism improves to avoid obstacles, it does not enhance efficiency on slopes or low-friction sur-faces, thus leaving limitations on locomotion reliability across varied terrain types.

Together, these innovations represent significant progress such as increased surface traction (Lee et al.), modular structural adaptability through different locomotion (Marin and Vivas), and movement over large obstacles (Fu and Li). Each addresses a different part of the terrain problem, but the combination of several approaches as with a passive mechanical system

can create an ideal solu-tion. Snake robots still lack an integrated mechanical system capable of simultaneously generating reliable grip, limit torque usage, and stability across unpredictable terrains such as rubble piles, steep inclines, and mixed-surface environments. This gap underlines the need for a new adaptive gripping mechanism that features traction, stability, and the accurate replication of real-life snakes.

Brainstorming the design

To develop a traction mechanism system for a snake robot, I explored multiple design concepts before narrowing to my proposed solution. The criteria to propose a mechanical design that adds to the original snake robot design that replicate snake-like motion, and a system that would remain mostly passive to minimize energy usage on the robot.

My initial idea was the use of tiny legs or spikes, similar to how spiders grip rough terrain. These tiny legs could dig into dirt, rock and be deployed only when additional grip was required. While this concept offered strong traction on irregular surfaces, it was ultimately too complex. It relied on many actuators and motors, increasing both energy con-sumption and mechanical difficulty. Additionally, it would not perform well on smooth or slippery surfaces. Overall, this idea wouldn't efficiently mimic real-life snakes' method for providing traction on different surfaces.

Another possibility was pressurized pads that would deploy to press against the ground and increase friction. This design would provide strong grip on inclines and uneven terrain but required a hydraulic or a complex pressurization system. These systems would add significant weight, slow down movement, and consume high amounts of energy, making it inefficient for a lightweight robot.

Another idea is a passive claw mechanism, consisting of small rubberized claws with textured surfaces. These claws could grip automatically on rough or inclined terrain and fold flat on smooth surfaces to minimize drag. To improve adaptability, I also considered incorporating smart materials, such as shape-memory alloys, that could react to terrain conditions to improve grip. Despite its advantages, this idea raised concerns of potential inconsistencies, and complexity when it does not effectively respond to external forces.

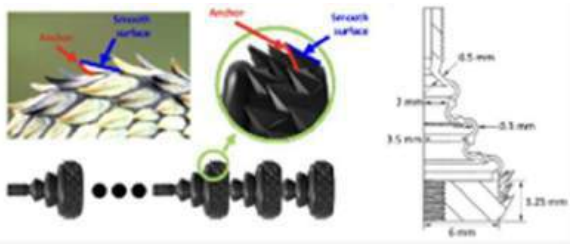
Through this process of exploration and evaluating different proposed ideas, I was able to identify the strengths and weak-nesses of different approaches, helping create the development of my final design that incorporates aspects of my original proposed ideas, along with a deployment system that can grip accordingly.

The Proposed Design

To address the limitations of grip, torque, and stability on steep or uneven terrain, this paper proposes a passive system that combines frictional mechanisms with a fully passive deployment method. This passive design will limit additional force or torque to function. The design functions as an add-on module to the underside of the snake robot, improving traction without interfering with its natural snake-like locomotion. This system will increase the ground

contact area, and use flexible brush-based materials with a full passive deployment method on the necessary terrain.

The foundation of the system is a rubberized silicon base with anisotropic frictional scales that replicate the directional resistance of snakeskin (Lee et al.). These scales allow high traction in the forward direction while minimizing drag or slippage during backward movement. The scales will have a smooth front surface that grips the floor. When the snake robot contracts the ground the front anchor grips while the smooth rear slides forward. This creates a forward movement by generating different friction forces in the opposite direction to prevent slippage (Lee et al.). This full base of the passive design is positioned along the underside of each snake robot's joint. It is curved outwards in order to increase surface area and the contact with the ground, helping the robot stabilize itself when climbing inclines and preventing it from slipping. This aspect of the proposed design is passive and ensures that more of the robot's body is actively engaged with the environment, reducing slip and limiting torque demands on each joint.



l scales on the rubberized base, showing directional resistance patterns that replicate snakeskin behavior for enhanced forward traction (Lee et al.).

At the center of each module of the attachment, there is a deployment of brushes that provides an additional adaptive gripping surface necessary to travel through rough terrain. The brushes are made of flexible, durable fibers that can naturally conform to terrain irregularities such as cracks, bumps, or rocks. On rough surfaces, the fibers spread out to stabilize against obstacles, while on smooth terrain they compress and fold back, reducing unnecessary drag. Their orientation can be adjusted based on the robot's direction of motion. For example, the brushes are angled forward when climbing, angled backward when moving downhill, and flattened on flat grounds. This passive adaptability makes the brush system highly versatile across different terrain conditions.

In order to adjust the brush system to adapt to the terrain and direction, this design integrates a small deployment system and retracts the brushes as needed. The deployment system uses a compact sliding panel that rotates the brush array downward into contact with the ground. When activated, it will push the sliding panel along a guided track, causing the brush assembly to pivot outward and downward. The brush system will automatically deploy based on the terrain and it will adapt during any other sh.ft.

By combining a rubberized anisotropic base, passive brush arrays, and a compact passive deployment system, this solution creates a modular gripping mechanism that

significantly en-hances snake robot locomotion. The design improves traction

Fig. 2. Deployment mechanism that folds outwards at uneven terrain by gravity

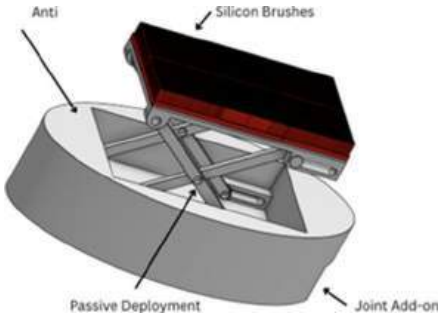


Fig. 3. Complete system overview of the adaptive gripping mechanism, showing the integration of the rubberized anisotropic base and deployable brush arrays.

On a variety of terrains such as inclines, it prevents slippage on smooth or loose terrain, and maintains efficient movement without requiring any increase in energy. The modular nature of this passive system allows it to be easily adapted or mounted to existing snake robot designs and thus extends their applicability to real-world tasks such as search-and-rescue operations or exploration in unpredictable environments (Han et al.). Overall, this system addresses the critical limitations of grip and stability while preserving the efficiency and flexibility that make snake robots unique.

Results and Analysis

The proposed mechanism is expected to address the traction issues mentioned earlier in the paper with soft and modular snake robots (Lee et al.). Instead of depending solely on surface-dependent friction features, such as anisotropic scales that improve grip only under certain conditions, the new system increases normal force to the ground points at instants in heavy elevation and frequent slippage through passive mechanical ac-tion. Because the mechanism does not rely on surface friction, it will maintain consistent locomotion on smooth glass, rough terrain, and mixed-frictional environments, conditions where earlier designs showed significant difficulties in performance.

In comparison with the 12-DOF modular design that im-proved heading accuracy primarily through gait-based control strategies (Marin and Vivas), the present design enhances directional stability by preventing lateral drift at its source. The modular robot's improvement from a heading error of roughly -7 degree to about -1 degree required continuous adjustments

based on terrain feedback (Transeth et al.; Marin and Vivas). The new system stabilizes each segment through increased traction, reducing the likelihood that small obstacles will redirect the robot's head. This decreases reliance on adjustment control sequences and allows more consistent alignment across unpredictable terrain without depending on complex corrective algorithms.

The system preserves the advantages of the advanced snake robot locomotion and movement algorithms for navigating rough environments while simultaneously increasing the force exerted to the point on the ground to maintain trac-tion during incline climbing (Fu and Li). This multipli-functional passive method will eliminate the trade-off seen in earlier designs, enabling stable performance across challenging terrain com-binations such as rubble piles, inclined surfaces, and regions with abrupt shifts in friction.

Conclusion

Overall, through this paper it highlights how existing snake robot designs each solve only part of the terrain problem, leaving major gaps in traction, reducing energy requirement, and stability when moving across steep or unpredictable envi-ronments. The anisotropic scale systems improve directional grip on smooth surfaces but do not increase normal force (Lee et al.). Modular designs with advanced movement improve heading accuracy but still depend heavily on control strategies to compensate for limited griping (Marin and Vivas). These findings emphasize the need for a more complete mechanical solution that supports consistent movement across a wide range of terrain types. The proposed passive system combines a rubberized anisotropic base, adaptive brush arrays, and a passive deployment panel, elements that work effectively with previous designs, and offers a theoretical approach for increasing traction, reduc-ing slippage, and improving stability without compromising snake-like motion. This design suggests a new direction for future snake robot development. By integrating passive friction enhancement with selective active deployment, the mechanism provides both adaptability and efficiency while addressing the key limitations observed across previous designs. Future steps could involve prototyping scaled versions of the mechanism, testing how the brush arrays behave across real materials. Additional research could also explore durability, energy consumption, and long-term wear of the components under real-world conditions. These next steps would help determine how well this passive system concept can be implemented into practical applications in environments such as the medical (Wang et al.) and exploration fields (Han et al.; Liu et al.), where reliable traction and stability are essential.

Acknowledgement

I would like to express my appreciation to Reza Oftadeh for his mentorship and guidance through this research paper. His expertise in robotics and mechanical systems was very valuable for the research process and brainstorming the final proposed passive gripping design. His constructive feedback and guidance contributed to the final paper.

Works Cited

Fu, Qiyuan, and Chen Li. “Robotic Modelling of Snake Traversing Large, Smooth Obstacles Reveals Stability Benefits of Body Compliance.” Royal Society Open Science, vol. 7, no. 2, Feb. 2020, p. 191192. https://doi.org/10.1098/rsos.191192.

Han, Sangchul, et al. “Snake Robot Gripper Module for Search and Rescue in Narrow Spaces.” IEEE Robotics and Automation Letters, vol. 7, no. 2, Apr. 2022, pp. 1667–73. https://doi.org/10.1109/LRA.2022.3140812.

Hirose, Shigeo, and Hiroya Yamada. “Snake-like Robots [Tutorial].” IEEE Robotics & Automation Magazine, vol. 16, no. 1, 2009, pp. 88–98, https://doi.org/10.1109/MRA.2009.932130.

Lee, Seonghyeon, et al. “Snakeskin-Inspired 3D Printable Soft Robot Composed of Multi-Modular Vacuum-Powered Actuators.” Actuators, vol. 12, no. 2, Feb. 2023, p. 62. https://doi.org/10.3390/act12020062.

Liu, Jindong, et al. “Review of Snake Robots in Constrained Environments.” Robotics and Autonomous Systems, vol. 141, July 2021, p. 103785. https://doi.org/10.1016/j.robot.2021.103785.

Marin, Jairo J., and Oscar A. Vivas. “Design and Evaluation of a Snake Robot with Bio-Inspired Locomotion over Uneven Terrain.” International Journal of Mechanical Engineering and Robotics Research, vol. 14, no. 1, 2025, pp. 78–86. https://doi.org/10.18178/ijmerr.14.1.78-86.

Transeth, Aksel Andreas, et al. “Snake Robot Obstacle-Aided Locomotion: Modeling, Simulations, and Experiments.” IEEE Transactions on Robotics, vol. 24, no. 1, Feb. 2008, pp. 88–104. https://doi.org/10.1109/TRO.2007.914849.

Wang, Honghong, et al. “Continuum Robots and Magnetic Soft Robots: From Models to Interdisciplinary Challenges for Medical Applications.” Micromachines, vol. 15, no. 3, Mar. 2024, p. 313. https://doi.org/10.3390/mi15030313.

Can Music Reduce Stress in Adolescents? A Systematic Review of Music Listening and Cortisol Regulation By Prathiksai Amathi

Abstract

Adolescent stress represents a public health concern with documented neurobiological consequences for the developing brain. This literature review seeks to examine the extent to which music listening influences cortisol levels and stress regulation. This is specifically within adolescent populations. This review uses 11 peer-reviewed sources such as controlled experiments, scoping reviews, meta-analyses, and editorial commentaries published between 2013 and 2025. The sources were found through searches of PubMed, Frontiers, Google Scholar, and JMIR databases. The findings were organized across four thematic areas: evidence that music reduces cortisol, the moderating factors that influence the magnitude of this effect, adolescent-specific findings, and mixed or contradictory findings.

This review finds that music listening may reduce both perceived stress and physiological stress markers such as cortisol. The magnitude grows when music is self-selected, slow-tempo, and instrumental and administered after the initial stress response. Nonclinical populations demonstrate stronger responses to music interventions than clinical populations. Limited adolescent-specific findings also suggest that this population is both vulnerable to chronic stress and potentially highly responsive to music-based interventions. Methodological inconsistencies across studies and the scarcity of adolescent-targeted research represent the most significant gap in the current literature. Therefore, future research should aim to prioritize adolescent-targeted intervention studies using evidence-based musical parameters in naturalistic settings.

Keywords Salivary cortisol, adolescent stress, music listening, HPA axis, stress intervention, perceived stress,

Introduction

Stress in adolescents has proved to be an enormous problem due to current research demonstrating that teenagers experience stress at higher rates than those of adults (American Psychological Association). The consequences of this stress have been documented. When the hypothalamic-pituitary-adrenal (HPA) axis is activated frequently and for long periods of time, immune function, memory consolidation, and mood regulation are all disrupted and destabilized (Vadnie et al.). Along with this, research indicates that stress exposure during this developmental window for adolescents can produce neurobiological changes in the circuits including reward, anxiety, and executive functioning (Vadnie et al.).

Despite how large this problem is, interventions have remained inaccessible or flawed for many adolescents suffering from prolonged stress exposure. So, this problem creates a need for evidence-based, non-pharmacological stress reduction strategies that adolescents can easily access. Music listening has emerged as a promising stress reduction tool due to its accessibility,

low cost, and support by an increasing body of research. In light of this, this review examines such evidence, paying close attention to findings relevant to adolescent populations.

This review aims to answer the question: to what extent does music listening influence cortisol levels and stress regulation in adolescents? Currently, evidence suggests that music listening may reduce both perceived stress and physiological stress markers, however, the magnitude of the effects varies based on music type, individual preference, timing, and study design. To evaluate this evidence, this review examines findings across four thematic areas: the evidence that music reduces cortisol, the factors that influence the magnitude of this effect, adolescent-specific findings, and contradictory or mixed findings. These themes provide an overall complex picture of music's potential as an evidence-based stress intervention for adolescents.

Background

As soon as a threat occurs, the hypothalamic-pituitary-adrenal (HPA) axis is activated. When this happens, the hypothalamus activates the pituitary gland, causing the adrenal glands to release cortisol straight into the bloodstream. Cortisol can be useful in short bursts because it sharpens focus and prepares the body for response. However, when stress is too frequent and the HPA axis is activated tirelessly without adequate recovery time, cortisol elevation can be highly dangerous and over time, the effects can include weakened immune function, memory, and mood regulation (Vadnie et al.). The prefrontal cortex, which governs decision-making and emotional regulation, does not fully develop until the mid-twenties. Chronic stress during this window causes neurobiological damage.

Research using animal models has demonstrated that circadian rhythm disruption during adolescence significantly alters dopamine signaling in the nucleus accumbens and prefrontal cortex, which increases risk-taking behavior, harms spatial working memory, and disrupts anxiety responses (Vadnie et al.). Researchers have looked to music as a potential intervention because when someone listens to music, the mesolimbic reward pathway lights up, triggering the release of dopamine, oxytocin, and serotonin, all of which play important roles in stress reduction and emotional stability (Zhang et al.). Researchers have relied heavily on salivary cortisol, which is a non-invasive and reliable biomarker of HPA axis activity to measure if music actually affects stress physiologically. Salivary cortisol is the primary stress marker and it is what most studies in this review measured.

Methodology

To find the right studies for this review, searches were run across PubMed, Frontiers, Google Scholar, and JMIR using search terms like 'music cortisol', 'music stress adolescents', 'salivary cortisol music intervention', and 'HPA axis music'. From those searches, 11 papers were used. The bar for inclusion was that a study had to actually measure cortisol or perceived stress as a primary outcome and use music as the main intervention. The sources range from 2013 to 2025 and include a mix of study types, from controlled experiments and meta-analyses to

scoping reviews and editorial commentaries. Studies were left out if they didn't report cortisol or stress outcomes, or if they focused exclusively on clinical populations with diagnosed neurological or psychiatric conditions, as the goal of this review is to understand music's effects on generally healthy adolescents, not those with pre-existing conditions.

Evidence That Music Reduces Cortisol

Among the most methodologically sound studies in this review, Thoma et al. took 60 healthy female volunteers with a mean age of 25 and randomly split them into three groups before putting them through a standardized psychosocial stress test called the Trier Social Stress Test (TSST), which is a structured situation designed to induce acute stress. One group listened to relaxing music, another to rippling water, and the last group sat in silence. The group that listened to the relaxing music actually had the highest cortisol levels during the stressor. But once it was over, that same group recovered back to baseline faster than either of the other two. This pattern suggests that music may function more effectively as a tool in cortisol level recovery rather than cortisol level prevention. This timing distinction carries important implications for intervention design, as addressed in the moderating factors section of this review. Importantly, the music selected was experimenter-selected rather than self-chosen, a limitation that likely produced an underestimate of music's true potential in real world stress reduction situations.

A review of 34 studies spanning from 1990 to 2024 found that music listening, particularly self-selected and classical pieces, consistently brought down physiological stress markers like cortisol. The music also reduced heart rate variability and blood pressure across a wide range of study designs and populations (Saskovets et al.). The review also identified that natural sounds such as running water demonstrated potential in stress reduction, though the evidence still remains small and less conclusive than music. The findings across 34 studies and three decades of research provide strong and broad support for music's capacity to attenuate physiological stress responses.

Likewise, a 2025 meta-analysis that examined studies across Web of Science, Scopus, and PubMed, found that music-based interventions reduced cortisol across both clinical and nonclinical populations (Zhang et al.). Within this study, it was found that healthy individuals showed stronger responses to music than those with diagnosed conditions. This is most likely because their stress systems are intact and more responsive to outside stimuli. This is especially relevant to the present review, which focuses on adolescent populations without diagnosed conditions.

Moderating factors

While the preceding evidence suggests that music as a whole reduces cortisol, there are many complications that influence the magnitude. The key moderating factors are music type, individual preference, and the timing of the intervention. Chief among these factors is individual preference. Helsing et al. conducted a two-week field experiment where 21 women listened to self-selected music every day for 30 minutes while in their own homes. For one week the women

were asked to listen to relaxing music, and the second week they listened to energizing music as their cortisol was measured across multiple collection points. Both music conditions produced significant stress reduction compared to control groups where participants listened to no music. The use of self-selected music in a real world setting produced far greater results than experimenter-controlled music in controlled laboratory studies such as Thoma et al. This suggests the core effect is how music exerts its effects of stress reduction. In other words, the stress reduction does not come from the music alone, but the actual relationship between the music and the listener.

Beyond individual preference, the type of music and tempo of music selected significantly moderates its stress-reducing effects. Research examining music's role in academic stress reduction among young adults found that music with a slow tempo (particularly lo-fi and classical music) was associated with cortisol reduction and calming effects. High-tempo music like rock or electronic, on the other hand, actually worked against stress reduction (Khushboo and Mishra). Lyrical music was more complicated, suggesting that during studying or other academic tasks, instrumental music is generally a better choice. Moreover, these findings reinforce the conclusion that individual music preference remains the strongest determinant of stress reduction outcomes. The timing of music listening represents a third critical moderating factor. Evidence suggests that music does not function uniformly across all points of the stress cycle. Thoma et al. found that music listening was a significantly more effective post-stress recovery tool rather than a stress prevention tool. Participants that listened to music before a stress response showed faster cortisol recovery afterward. This distinction has important implications for how music-based interventions should be administered.

Likewise, Helsing et al. demonstrated that daily music listening over two weeks produced stress reduction in a real-world setting. These findings suggest that music functions most effectively as a stress reduction mechanism when administered consistently over time and after stress exposure. Taken together, these moderating factors suggest that music's effect on cortisol is not uniform but highly dependent on context. An effective music-based stress intervention for adolescents would need to account for individual preference while also favoring slow-tempo instrumental music administered over consistent periods of time following stress exposure rather than before it. Importantly, the evidence suggests that no single piece of music will produce identical effects across all listeners, and that intervention design must treat preference, type, and timing not as secondary considerations but as core determinants of efficacy.

Adolescent Specific Findings

During this period, the adolescent brain undergoes significant and structural changes as it begins to develop. This causes it to be vulnerable to the effects of chronic stress and sleep disruptions. Research also finds that during this window of development, many mood and anxiety disorders typically begin to take place. This demonstrates how critical this developmental window is and how stress can have permanent effects. However, very few studies have been conducted to examine music listening as a stress reduction tool for adolescents

specifically. An editorial which synthesized current research on adolescent stress emphasized the fact that sleep disruptions and chronic stress are documented risk factors for the development of mood and anxiety disorders (Vadnie et al.). This editorial raises a very specific question: can these effects be reversed or are they permanent?

This question has direct implications for the development of accessible interventions (including music-based approaches) for adolescent populations. Experimental evidence using animal models begins to answer this question. DePoy et al. exposed adolescent mice to circadian rhythm disruption and measured behavioral consequences in their adulthood. These results demonstrated lasting effects including heightened sensitivity to reward, increased risk-taking behavior, and anxiety-like behavior. However, when the adult mice underwent the same disruption, these effects were not observed. This finding directly suggests that the adolescent brain is not simply a smaller version of the adult brain, but a distinct developmental system in which stress exposure produces uniquely lasting neurobiological consequences. Intervening during adolescence rather than adulthood may be critical for preventing long-lasting neurobiological damage that chronic stress produces.

Among the few studies to actually examine music listening specifically in adolescent populations, Chen et al. investigated temporal factors of music listening on stress reduction in 252 middle and high school students, and the findings demonstrated that music listening reduced perceived stress in this school-aged population, providing direct empirical evidence for music's stress-reduction potential in adolescents rather than adults. Taken together, the evidence reviewed in this section suggests that adolescents represent both the population most vulnerable to chronic stress and the population most in need of accessible interventions. The relative scarcity of adolescent-specific music and cortisol research represents a significant gap in the literature that needs to be prioritized.

Contradictory and Mixed Findings

Despite the evidence in the preceding section, the literature contains important information that contradicts and complicates the universal claims about music's capacity to reduce cortisol. A pilot study examining music listening therapy in children and adolescents with autism spectrum disorder (ASD) found no significant cortisol reduction following 30 days of classical music intervention (Florezi et al.). This finding stands in contrast to the effects observed in the nonclinical populations reviewed earlier. This evidence is also consistent with the meta-analytic finding that clinical populations demonstrate weaker responses to music interventions than healthy individuals (Zhang et al.). This finding highlights the fact that music's stress reduction effects are dependent on the population it's affecting and cannot be generalized across all groups without consideration of individual neurological differences. Likewise, a 2025 randomized controlled trial protocol examining music with auditory beat stimulation on anxiety in students with trait anxiety acknowledges that existing studies differ considerably in music type, intervention duration, cortisol sampling windows, and population characteristics, making direct comparisons across studies difficult (Ueberholz et al.). The fact that a rigorous RCT is still

being designed demonstrates just how much methodological work remains to be done before a definitive conclusion can be drawn.

Taken together, these contradictory and mixed findings do not undermine the evidence reviewed in Themes 1 through 3; rather they give a broader context. Music's effect on cortisol is conditional; it depends on population, music type, individual preference, and methodological design. A literature that acknowledged only supportive findings would be incomplete and scientifically irresponsible. The contradictions reviewed here strengthen rather than weaken the case for rigorous adolescent-specific research into music as a stress intervention.

Discussion

The evidence reviewed across four thematic areas converges on a consistent conclusion: music listening influences cortisol and perceived stress in meaningful ways, though the magnitude and reliability of this effect depend substantially on various moderating factors, including individual preference, music type, timing, and population characteristics. Themes 1 through 3 establish that music reliably reduces salivary cortisol in nonclinical populations when administered after the initial stress response and using self-selected music or slow and instrumental music over consistent periods of time. In contrast, Theme 4 complicates this picture by demonstrating that clinical populations respond differently, and that methodological inconsistencies across the literature make definitive universal claims premature.

The most significant gap shown by this review is the scarcity of research examining music as a cortisol intervention specifically in adolescent populations. The majority of studies reviewed here used adult participants, limiting the generalizability of findings to teenagers (which was established as the population most neurobiologically vulnerable to chronic stress) and least studied in this context. Research has consistently demonstrated that stress exposure during adolescence produces lasting neurobiological changes (DePoy et al.), suggesting that the adolescent brain is not simply a younger version of the adult brain but a developmentally unique system with unique responses to both stress and potential interventions. Among the few studies to examine music specifically in adolescent populations, Chen et al. demonstrated stress reduction effects in 252 middle and high school students. This evidence remains an isolated finding. A targeted program of research examining music listening as a cortisol intervention specifically in healthy adolescent populations, using self-selected or parameter-based music stimuli administered consistently following stress exposure, represents the most pressing direction for future work in this area.

Based on the findings reviewed here, an evidence-based music intervention targeting cortisol reduction in adolescents would incorporate several key design principles. To begin, music selection should prioritize individual preference rather than standardized stimuli, as the evidence generally demonstrates that personally resonant music produces stronger stress-reducing effects than experimenter-selected pieces (Helsing et al.). Secondly, the musical characteristics of selected pieces should favor slow tempo rather than fast tempo. Evidence shows that approximately 60 to 80 beats per minute instrumental pieces with consonant

harmonic structure are most consistently associated with stress reduction (Khushboo and Mishra). Third, the intervention should be administered following stress exposure rather than before it, given that music appears to function more effectively as a recovery tool rather than a prevention mechanism (Thoma et al.). Finally, the intervention should be sustained over multiple weeks of consistent daily listening rather than delivered as a single session, as cumulative exposure appears to amplify physiological benefits (Helsing et al.). An intervention designed around these principles (targeting healthy adolescent populations in naturalistic settings such as schools) would represent a meaningful advance on the existing literature and a genuinely accessible and low-cost approach to one of adolescence's most pressing public health challenges.

Conclusion

This review set out to examine the extent to which music listening influences cortisol levels and stress regulation in adolescents. The evidence reviewed supports the conclusion that music listening may reduce both perceived stress and physiological stress markers; however, the outcome varies as music type, individual preference, timing, and design all have a considerable impact. Foundational studies along with syntheses of the literature converge on the finding that slow-tempo, self-selected, instrumental music administered following the initial stress response is the most effective use of music as a tool for cortisol reduction in nonclinical populations. Research specifically examining music as a stress intervention in adolescent populations remains scarce; however, the limited evidence suggests that this population is both uniquely vulnerable to chronic stress and shows promising responses to music-based interventions. Future research should prioritize rigorous, adolescent-targeted studies using music with evidence-based parameters in naturalistic school settings. Given what we now know about the lasting neurobiological consequences of adolescent stress, accessible and evidence-based interventions are a necessity.

Works Cited

American Psychological Association. Stress in America 2023: A Nation Recovering from Collective Trauma. APA, 2023, www.apa.org/news/press/releases/stress/2023/collective-trauma.pdf

Chen, Y., et al. "Temporal Factors of Listening to Music on Stress Reduction." Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2019, doi.org/10.1145/3341162.3346272

DePoy, L. M., et al. "Adolescent Circadian Rhythm Disruption Increases Reward and Risk-Taking." Frontiers in Neuroscience, vol. 18, 2024, p. 1478508, doi.org/10.3389/fnins.2024.1478508

Fiorezi, J. N., et al. "Effects of Music Listening Therapy on Reducing Cortisol Levels and DNA Damage in Children and Adolescents with Autism Spectrum Disorder: A Non-Pharmacological Pilot Study." Research in Autism, vol. 125, 2025, p. 202624, doi.org/10.1016/j.reia.2025.202624

Helsing, M., et al. "An Experimental Field Study of the Effects of Listening to Self-Selected Music on Emotions, Stress, and Cortisol Levels." Music and Medicine, vol. 8, no. 4, 2016, pp. 187-198

Khushboo, and R. Mishra. "Impact of Music on Stress and Academic Performance among College Students." International Journal of Creative Research Thoughts, vol. 13, no. 3, 2025, ijcrt.org/papers/IJCRT2503702.pdf

Saskovets, M., et al. "Effects of Sound Interventions on the Mental Stress Response in Adults: Scoping Review." JMIR Mental Health, vol. 12, 2025, p. e69120, doi.org/10.2196/69120

Thoma, M. V., et al. "The Effect of Music on the Human Stress Response." PLOS ONE, vol. 8, no. 8, 2013, p. e70156, doi.org/10.1371/journal.pone.0070156

Ueberholz, L., et al. "Music with Auditory Beat Stimulation for Anxiety in University Students: Protocol for a Randomized Controlled Trial." Frontiers in Psychology, 2025, pmc.ncbi.nlm.nih.gov/articles/PMC12182183/

Vadnie, C. A., et al. "Editorial: Long-Term Effects of Adolescent Stress, Sleep Deprivation, or Circadian Disruption on Mood and Anxiety." Frontiers in Neuroscience, vol. 19, 2025, p. 1633483, doi.org/10.3389/fnins.2025.1633483

Zhang, J., et al. "Impact of Music-Based Interventions on Subjective Well-Being: A Meta-Analysis of Listening, Training, and Therapy in Clinical and Nonclinical Populations." Frontiers in Psychology, vol. 16, 2025, p. 1608508, doi.org/10.3389/fpsyg.2025.1608508

From Sharks to Humans: Comparative Tooth Regeneration Across Vertebrates and Their Implications for Human Regenerative Dentistry By Kate Kim

Abstract

Teeth serve crucial functions in mastication, speech, aesthetics and quality of life. Existing treatments such as dental implants, bridges and dentures can restore oral function, but cannot fully replicate the biological properties of natural teeth. Hence, regenerative dentistry has emerged as a promising field with the goal of regenerating functional teeth using stem cells, signaling molecules and biomaterial scaffolds. But complete biological tooth regeneration is still a great problem.

This paper compares the mechanisms of tooth regeneration across vertebrates, including sharks, fishes, amphibians, reptiles, and mammals, and compares these mechanisms with human tooth development. While humans lose the ability to regenerate teeth after permanent dentition due to degeneration of the dental lamina, many polyphyodont vertebrates maintain a persistent dental lamina that continuously supplies progenitor cells for lifelong tooth replacement. Although there are differences in the tooth replacement pattern among species, many regenerative mechanisms including epithelial-mesenchymal interactions and conserved signaling pathways such as Wnt, BMP, SHH, FGF and Sox2 are shared throughout vertebrate evolution.

Potential clinical applications including current regenerative strategies such as induced pluripotent stem cells (iPSCs), adult stem cells, signaling molecules, biomaterial scaffolds and anti-USAG1 therapy are also explained. This review suggests, on the basis of comparative evidence, that current regenerative dentistry is limited not only by the absence of stem cells, but by the missing developmental environment provided by the dental lamina. Reconstructing or reactivating this developmental niche may be a promising direction for future whole-tooth regeneration. Comparative studies of naturally regenerating vertebrates thus provide useful information for developing biological strategies that are capable of restoring fully functional human teeth.

Key Words

Dental regeneration, Stem cells, dental lamina, Polyphyodont, Monophyodont, Diphyodont, Regenerative dentistry

Introduction

People use their teeth every day and do not think much about them. They take their teeth for granted. But teeth are a key part of a person's overall health and well-being. They basically chew the food into smaller bits to make it easier to swallow and more efficient in digestion. In many animals, teeth are also important for catching prey or protecting against predators. In contrast, human teeth have other functions besides nutrition. They help to speak clearly, keep the natural shape of the face, influence appearance and self-confidence. Because of these reasons, it is important to have healthy teeth not only to function well orally, but also for the overall quality of life of a person (Chen and Liu).

Humans normally have only two sets of teeth in their lifetime, but many vertebrates replace their teeth continuously. The first set of teeth, called primary (deciduous) teeth, begins to emerge in early childhood and is later replaced by permanent teeth. Once the permanent dentition is fully developed, there generally is no further natural replacement. However, these teeth may be lost due to conditions such as dental caries, periodontal disease, trauma or developmental abnormalities. Losing teeth can affect one's ability to chew properly, speech, appearance and general health, so developing reliable ways to restore function and beauty is an important goal of contemporary dentistry (Chen and Liu).

There are three main ways to replace missing teeth – dental implants, bridges or dentures. Although these treatments can improve the ability to chew and restore the appearance of the mouth, they are still artificial replacements and cannot function fully like natural teeth. They do not completely replace the natural support and sensory function of the periodontal ligament and may also develop complications with time. Because of these limitations, researchers are developing new regenerated approaches to replace missing teeth in a more natural and effective way (Chen and Liu; Sui et al.).

Regenerative dentistry is an emerging field that combines dentistry, biology, and tissue engineering to develop novel approaches to the treatment of damaged or lost oral tissues. Its goal is to regenerate natural tissues by using the healing capacity of the body itself. Most regenerative approaches are based on three key components, including stem cells, biological signaling molecules, and biomaterial scaffolds supporting tissue growth. There has been great progress, but regeneration of fully functional teeth and supporting tissues remains a major scientific challenge (Jansen; Morita et al.).

Methodology

This literature review was conducted using comprehensive electronic databases, including PubMed, ScienceDirect, and Google Scholar, to identify published articles in English between 2013 and 2025. Key search terms included "tooth regeneration," "vertebrate tooth replacement," "polyphyodont," "diphyodont," "dental lamina," "dental epithelial stem cells," "successional lamina," "regenerative dentistry."

Background

I. Human Tooth Development and the Loss of Regeneration Capacity

Human Tooth Development

Human tooth development begins when the oral epithelium thickens to form the dental lamina, which subsequently initiates the formation of tooth germs.The dental lamina secretes signaling molecules such as bone morphogenic protein (BMPs) and fibroblast growth factor (FGFs) into the surrounding mesenchymal tissue which triggers the development of a tooth germ. The development of the tooth germ takes place in three main stages. These are the bud stage, cap stage and bell stage. The size and shape of the developing tooth are controlled by a

temporary signalling center known as the enamel knot during the cap stage. Different tissues start to specialize at the bell stage. These tooth tissues have two main embryonic origins. The oral ectoderm forms the enamel organ, which will eventually form the enamel. The dental mesenchyme differentiates into the dental papilla, the precursor of dentin and dental pulp. In addition, the dental follicle forms the periodontal ligament and the alveolar bone. These coordinated epithelium-mesenchymal interactions are necessary for the formation of a functional tooth (Morita et al.).

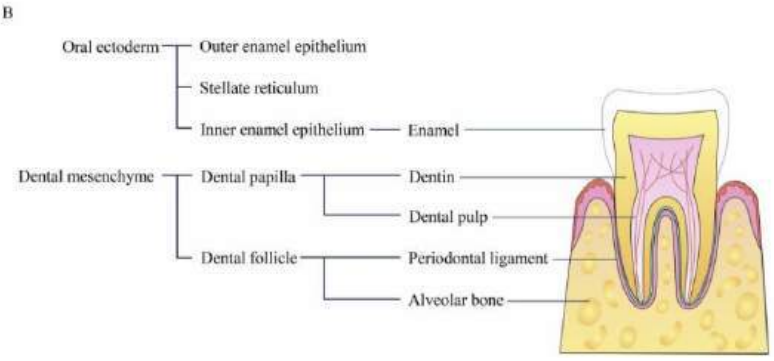


Figure 1. Human Tooth Development and the Embryonic Origin of Major Dental Tissues
Reproduced from Morita et al. (2025), CC BY-NC-ND 4.0

Why Humans Cannot Naturally Regenerate Teeth

Unlike most other vertebrates, humans can replace their teeth only once in their lifetime. Humans are diphyodont, that is, they develop two generations of teeth, primary (deciduous) teeth, followed by permanent teeth. As the permanent dentition develops, the dental lamina degenerates progressively by programmed cell death (apoptosis), leaving small epithelial remnants, the dental lamina rests (DLRs). These remainders have long been regarded as biologically inactive, which explains why humans cannot produce additional teeth naturally once the permanent dentition is formed. However, recent studies have shown that DLRs still express several stem cell markers indicating that humans may have a limited regenerative potential (Fraser et al., “Shark Tooth Regeneration”).

II. Comparative tooth replacement system among vertebrates

Natural Tooth Regeneration in Animals

Interestingly, animals like sharks and crocodiles can naturally regrow their teeth throughout their lifetime. The way these animals restore their teeth can teach scientists how to develop new regenerative therapies. This review explores the biological mechanisms of tooth regeneration in animals, compares them with those involved in human tooth development and discusses implications for future regenerative dentistry.

Regeneration of teeth is highly variable among vertebrates. The number of tooth replacement cycles has generally decreased over the course of vertebrate evolution, a trend

believed to be related with the evolution of specialized diets and the need for specific occlusion in mammals. Most fish are polyphyodonts, continuously replacing their teeth throughout their lives. Humans and most other mammals are diphyodont, with only two sets of teeth, while rodents and other monophyodont mammals have only one. But there are exceptions, such as elephants and manatees, that can replace their teeth several times (Tucker and Fraser).

The formation and regeneration of teeth require a specialized epithelial structure, the dental lamina, that arises early in tooth development. In polyphyodont animals, the dental lamina is active throughout life and constantly generates replacement teeth. Conversely, in mono- and diphyodont species, the dental lamina degenerates progressively, preventing the development of new tooth germs and greatly restricting the ability to regenerate further teeth (Fraser et al., “Shark Tooth Regeneration”).

Main Strategies for Tooth Replacement

Tooth development is controlled by highly conserved biological processes but the ability to replace teeth differs widely between vertebrates. Sharks, crocodiles, fishes, reptiles and mammals differ greatly in their tooth shape, rate of replacement and feeding strategies. However, they share many of the same developmental signaling pathways and epithelial-mesenchymal interactions. The evolutionary conservation implies that research on naturally regenerating animals may offer crucial insights into human tooth regeneration. Instead of having to find totally new biological mechanisms, regenerative dentistry might be able to reactivate developmental programs already present in humans (Tucker and Fraser).

There are two main ways to replace teeth, the one-for-one replacement system and the many-for-one replacement system. In the one-for-one replacement system each functional tooth is replaced by one successor tooth. This pattern is shared by many teleost fishes and crocodilians. The many-for-one replacement system is typical of sharks and many reptiles . In this system several generations of replacement teeth are formed at once before eruption . This creates a conveyor-belt-like dentition that allows for rapid and continuous tooth replacement throughout life without the loss of feeding function (Fraser et al., "Shark Tooth Regeneration").

Sharks: A Model for Continuous Tooth Regeneration

Sharks are one of the best animal models to study tooth regeneration because they can continuously replace their teeth throughout their life. Unlike humans, sharks possess a persistent dental lamina, a specialized epithelial structure that houses stem and progenitor cells and continuously generates new teeth. These stem cells make new tooth buds when old teeth fall out, effectively providing an infinite supply of replacement teeth. Studies have revealed that the molecular mechanisms regulating tooth regeneration are well conserved among vertebrates . Formation of the dental lamina and the maintenance of the new teeth are controlled by Sox2, Wnt/β-catenin and BMP, important signaling pathways and stem cell markers. These pathways are also used by humans during the development of early teeth. So, by studying sharks, we could learn why humans lose the ability to regenerate teeth once we get our permanent teeth.

In humans the dental lamina slowly degenerates after the formation of the permanent dentition while dental lamina rests (DLRs) remain. These remnants are generally quiescent, but continue to express many markers of stem cells, suggesting that they retain some potential for regeneration. In rare cases, DLRs can be reactivated and lead to the development of benign odontogenic tumours such as ameloblastomas. Interestingly, the Wnt/β-catenin signaling, which plays a central role in tooth regeneration of polyphyodont animals is associated with these pathological changes. The results of these studies suggest that knowledge of how the Wnt/β-catenin signaling is regulated may allow researchers to safely induce tooth regeneration in humans in the future (Fraser et al., “Shark Tooth Regeneration”).

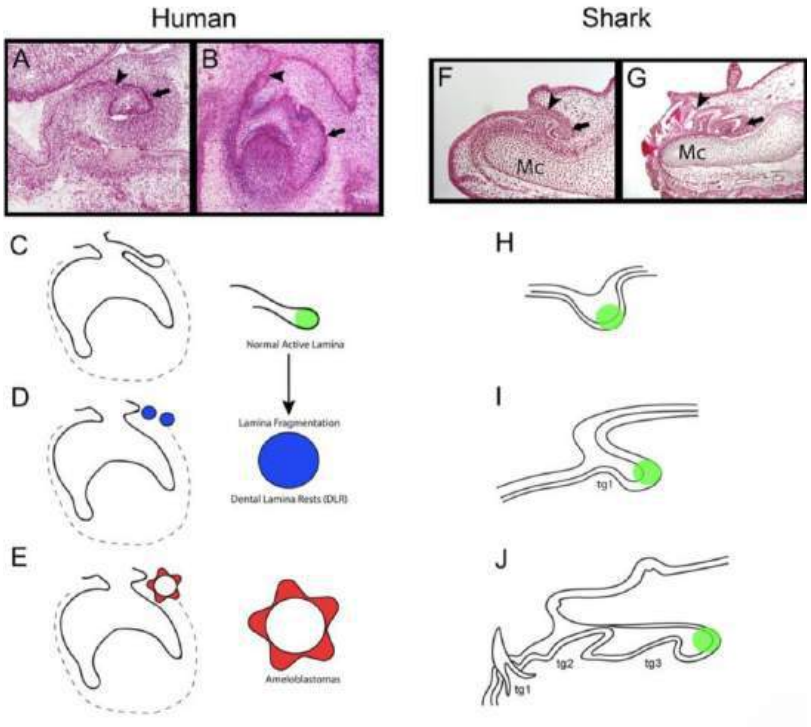


Figure 2. Comparison of Dental Lamina Fate in Humans and Sharks
In humans, dental lamina degenerates after permanent tooth development, leaving small epithelial remnants known as dental lamina rests (DLRs). In contrast, sharks maintain a persistent dental lamina that continuously supplies progenitor cells for lifelong tooth regeneration.
Reproduced from Fraser et al. (2020), CC BY 4.0.

Only sharks and rays have a rapid, conveyor-belt-like pattern of tooth replacement. In sharks all teeth are attached to a continuous dental lamina persisting throughout life. Sox2-positive stem cells continuously provide progenitor cells, which give rise to new tooth germs. Cyclic activation of Wnt/β-catenin signaling leads to the formation of successive replacement teeth, allowing for lifelong tooth regeneration without compromising feeding function (Fraser et al., “The Dental Lamina”).

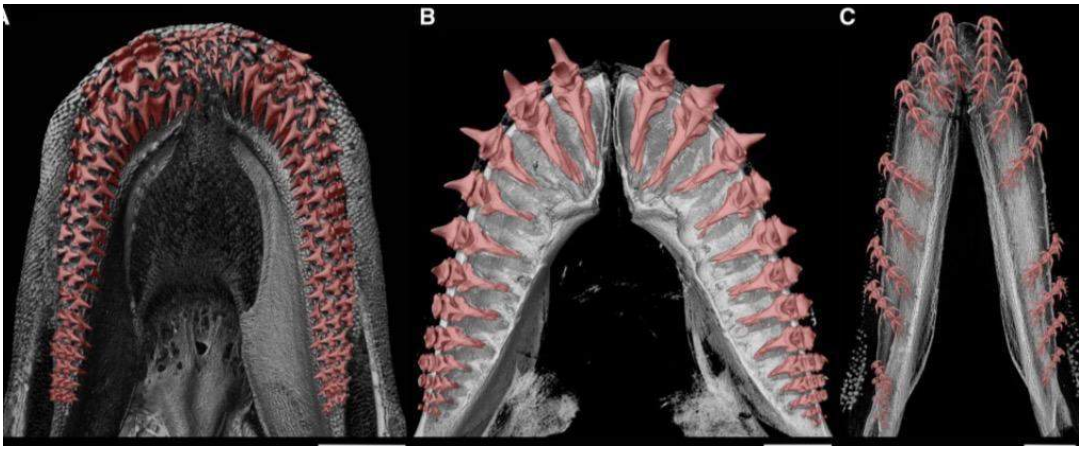


Figure 3. Micro-CT scans of three different sharks that show the diversity of dental patterning of dental spacing Reproduced from Fraser and Standing, "The Dental Lamina: An Essential Structure for Perpetual Tooth Regeneration in Sharks," Integrative and Comparative Biology (2020), by permission of Oxford University Press.

Tooth Regeneration in Fish

Most teleost fishes replace their teeth one for one. Sharks replace each functional tooth with one successor tooth for their whole life . Replacement teeth are formed by epithelial cells that come from the successional dental lamina and ensure that each new tooth grows in the place of the previous one. This replacement pattern is distinct from the conveyor-belt seen in sharks, but both strategies rely on the presence of an active epithelial source that can generate new tooth germs.

The continuous tooth replacement has also been a factor in the diversity of fish dentition. Cichlid fish, for example, develop more and more complex tooth shapes from generation to generation as teeth are replaced, and pufferfish develop from simple juvenile teeth to a specialized beak-like structure. These examples illustrate that tooth regeneration is important not only for replacing worn or damaged teeth, but also for driving morphological adaptation during growth and evolution (Tucker and Fraser).

Tooth Regeneration in Amphibians

Amphibians that are polyphyodonts, such as frogs and salamanders, have a constant cycle of tooth replacement throughout their lives, using a one-for-one replacement system. Like other polyphyodont vertebrates, they have a persistent dental lamina that connects the developing replacement teeth with the oral epithelium and provides epithelial cells for successive tooth formation. Tooth replacement is generally faster in juveniles than adults and may be affected by physiological changes accompanying growth and metamorphosis, e.g., thyroid hormones. The results indicate that amphibian tooth regeneration is regulated by not only local epithelial cells and developmental signaling pathways, but also by broader systemic factors. Thus, future regenerative therapies may need to consider both local tissue-engineering approaches and whole-body regulatory signals (Tucker and Fraser; Davit-Béal et al.).

Tooth Regeneration in Reptiles

Reptiles, including crocodiles, snakes and geckos, provide useful models for studying tooth regeneration. Polyphyodont reptiles, such as sharks, have a continual dental lamina that acts as a lifelong source of epithelial progenitor cells . Epithelial stem cells that produce replacement teeth are found in the successional lamina within this structure. In species of reptiles that do not replace their teeth, degeneration of the successional lamina prevents the initiation of new tooth germs, showing the essential role of the dental lamina in lifelong tooth regeneration. Interestingly, sharks always form new teeth irrespective of the loss of teeth, but in alligators the loss of a functional tooth can trigger the formation of a new tooth (Fraser et al., “The Dental Lamina”).

Tooth Regeneration in Mammals

Most mammals are diphyodonts, having only two sets of teeth in their lifetime. The dental lamina in mammals has an initial structure similar to that of other vertebrates, but after the establishment of the permanent dentition it degenerates to a large extent, preventing further tooth replacement. It is still not clear, however, if the loss of regenerative capacity is due to the disappearance of stem cells, or to the inability of the dental lamina to provide an appropriate microenvironment for the initiation of new tooth development. Interestingly, continuous tooth replacement still persists in some mammals such as elephants and manatees, suggesting that parts of the ancestral regenerative program have been evolutionarily conserved. These observations suggest that understanding the degeneration of the dental lamina in mammals may provide important insights for the future development of therapies for tooth regeneration (Tucker and Fraser; Fraser et al., “The Dental Lamina”).

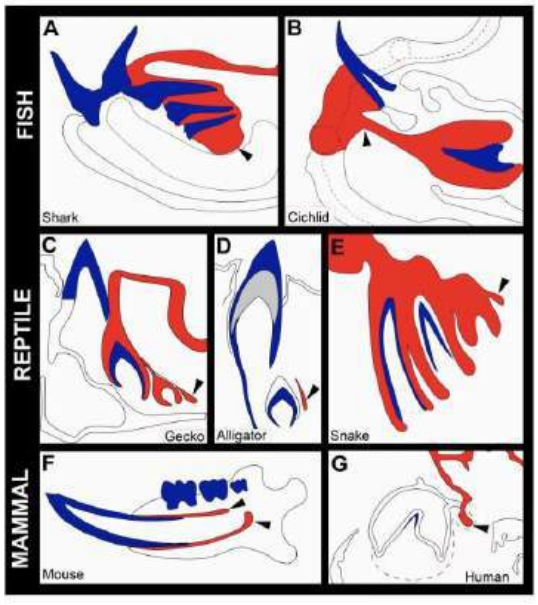


Figure 4. Comparison of Dental Lamina among Vertebrates
Reproduced from Fraser and Standing, "The Dental Lamina: An Essential Structure for Perpetual Tooth Regeneration in Sharks," Integrative and Comparative Biology (2020), by permission of Oxford University Press.

(A) Shark: many for one tooth replacement with a persistent dental lamina
(B) Cichlid: one for one tooth replacement supported by a persistent dental lamina
(C) Gecko: continuous polyphyodont replacement
(D) Alligator: many for one tooth replacement with persistent dental lamina
(E) Snake: many for one dental regeneration
(F) Mouse: monophyondont capable of continuous renewal of the incisor with no true replacement
(G) Human: Diphyodont dentition which regresses after permanent tooth development

III. Current Strategies for Human Tooth Regeneration

Induced Pluripotent Stem Cells (iPSCs)

Induced pluripotent stem cells (iPSCs) are adult somatic cells that have been genetically re-programmed to gain the properties of pluripotent stem cells. Similar to embryonic stem cells, iPSCs can differentiate into numerous cell types, such as dental epithelial and mesenchymal cells. iPSCs are a promising and personalized cell source for regenerative dentistry due to their ability to be derived from a patient's own cells.

Researchers have successfully differentiated iPSCs into several types of dental cells including odontoblasts, dental pulp cells, periodontal ligament (PDL) cells and ameloblast-like cells. In addition, tooth-like structures with enamel, dentin, pulp and cementum were generated, demonstrating the ability of iPSCs to recreate multiple components of a developing tooth. However, there are still remaining challenges to be overcome, including lowering the risk of tumorigenesis, long-term survival, functional integration and mechanical resilience of the regenerated teeth in the oral environment (Morita et al.; Sui et al.)

Adult Stem Cells

Adult stem cells (ASCs) are one of the most studied cell sources for regenerative dentistry, owing to their accessibility from tissues of a patient and a lower risk of tumor formation than embryonic stem cells. ASCs are naturally occurring in dental tissues and play a role in the maintenance and repair of the tissue, unlike induced pluripotent stem cells (iPSCs). Among the most promising sources are dental pulp stem cells (DPSCs) isolated from the dental pulp of permanent teeth, which are able to differentiate into odontoblast-like cells responsible for dentin formation and repair of the pulp-dentin complex. Another important source is stem cells from human exfoliated deciduous teeth (SHED) that are obtained from naturally shed primary teeth and have strong regeneration potential for both dental and bone tissues. Other populations of adult stem cells, including stem cells from the apical papilla (SCAPs) and periodontal ligament stem cells (PDLSCs), have also been shown to regenerate tooth roots, dental pulp and periodontal tissues.

Adult stem cells have a limited regenerative potential and are unable to create a whole tooth on their own, as a tooth is composed of several highly specialized tissues organized in a complex three-dimensional structure. So, for successful tooth regeneration, we need stem cells

and precise developmental signals. Tooth development is regulated by growth factors and cytokines controlling cell proliferation, differentiation and tissue organization. Bone morphogenetic proteins (BMPs) have been shown to promote the differentiation of dentin- and enamel-forming cells, whereas fibroblast growth factors (FGFs) and the Wnt signaling pathway are involved in the regulation of cell proliferation and lineage specification throughout tooth development (Morita et al.; Sui et al.).

Biological Signaling Molecules

Bone morphogenetic proteins (BMPs) promote the differentiation of stem cells into dentin- and enamel-forming cells. Fibroblast growth factors (FGFs) and the Wnt signaling pathway regulate cell proliferation, differentiation, and tissue patterning throughout tooth development. Researchers can manipulate the combination, concentration, and timing of these signaling molecules to improve stem cell differentiation and promote the formation of more organized dental tissues.

Growth factors are also widely used in tissue engineering. They can be incorporated into biomaterial scaffolds to stimulate stem cell recruitment, angiogenesis, and tissue regeneration at sites of injury. More recently, stem cell-derived exosomes, small extracellular vesicles secreted by stem cells, have emerged as a promising cell-free approach for delivering regenerative signals. Exosomes contain proteins, microRNAs, and other bioactive molecules that promote tissue repair while reducing some of the risks associated with stem cell transplantation.

Despite these advances, important challenges remain. Tooth development requires different signaling molecules at different developmental stages, meaning that their concentration, timing, and interactions must be precisely controlled. Further research is therefore needed to optimize these signaling pathways and improve the safety and effectiveness of regenerative therapies (Morita et al.; Sui et al.).

Biomaterial Scaffolds

Recent advances in scaffold fabrication techniques including electrospinning and 3D bioprinting have also enhanced scaffold design by producing structures closely mimicking the natural architecture of dental tissues. These technologies are capable of producing patient-specific scaffolds that enhance stem cell attachment, proliferation and tissue organization. (Morita et al.; Sui et al.) However, further optimization is needed to improve their long-term safety, mechanical stability and clinical applicability to be routinely applied in regenerative dentistry.

Biomaterial scaffolds are generally classified into natural and synthetic materials. Natural scaffolds such as collagen, fibrin and decellularized extracellular matrix (dECM) are highly biocompatible and mimic the native extracellular environment, thus facilitating cell attachment and differentiation. Conversely, synthetic scaffolds, such as polycaprolactone (PCL), poly(lactic-co-glycolic acid) (PLGA), and gelatin methacryloyl (GelMA), possess increased mechanical strength and control over degradation rate and scaffold architecture. An important

strategy in regenerative dentistry is the combination of the biological advantages of natural materials with the structural properties of synthetic materials (Morita et al.; Sui et al.).

Anti-USAG1 Therapy

USAG-1 (uterine sensitization-associated gene-1) acts as a natural inhibitor of BMP and Wnt signaling, which are two key pathways for tooth development. Researchers demonstrated that blockade of USAG-1 with monoclonal antibodies enhances BMP signaling and activates dormant tooth-forming potential. In mouse models, anti-USAG1 antibody therapy successfully induced formation of extra functional teeth without the need for stem cell transplantation or tissue engineering. More recently, this strategy has been moving towards clinical application. Anti-USAG1 therapy is in early clinical studies as a treatment for congenital tooth agenesis, a condition where patients are born without one or more permanent teeth. Although currently used only for certain developmental disorders, this method suggests that manipulation of developmental signaling pathways alone can induce the formation of new teeth. Future studies will examine the possibility of combining anti-USAG1 therapy with stem cells, biomaterial scaffolds, or artificial dental lamina to regenerate complete functional teeth (Murashima-Suginami et al.; Takahashi et al.).

Table 1. Summary of Key Studies on Vertebrate Tooth Regeneration and Their Clinical Implications

Authors	Year	Animal models	Methodology	Key findings	Potential Clinical Translation
Wu et al.	2013	Alligator	Histology, Tooth extraction, BrdU	Stem cell niche and the successional dental lamina identified. Tooth extraction initiated a new tooth replacement cycle	Injury induced regenerative signalling could potentially be activated to stimulate human tooth regeneration
Tucker and Fraser	2014	Comparative vertebrates including Chondrichthyans, Osteichthyans, Amphibians and reptiles, Amphibians and reptiles	Comparative developmental biology review	Demonstrated that vertebrates share highly conserved developmental pathways despite diverse tooth replacement strategies	Human tooth regeneration may require reactivating conserved developmental programs rather than creating new ones
Fraser et al.	2019	Human and shark	Histology, immunohistochemistry	Human dental lamina rests retain stem cell markers(e.g.,SOX2) similar to sharks, suggesting residual regenerative potential	Reactivation of dental lamina rests may become a future strategy for biological tooth regeneration

Fraser et al.	2020	Shark	Comparative anatomy and developmental biology	Persistent dental lamina continuously supplies epithelial progenitor cells throughout life, enabling lifelong tooth regeneration	Engineering artificial dental lamina or stem cell niche may enable continuous tooth regeneration in humans
Mura shima et al.	2024 - 2025	Mouse	Anti-USAG-1 antibody therapy	Blocking USAG-1 activated dormant tooth forming potential and induced new tooth formation	Demonstrates that manipulating developmental signaling can stimulate new tooth formation, supporting future regenerative therapies
Thiery et al.	2025	Shark	Transcriptomics and molecular profiling	Identified novel genes associated with lifelong maintenance of the shark dental lamina	Provides new molecular target for developing future regenerative therapies in human

Discussion

Dental lamina as the missing link in tooth regeneration

Current regenerative dentistry mainly emphasizes stem cells, signaling molecules and scaffolds. But studies on animals that naturally regenerate suggest that stem cells may not be enough. It has been shown that the dental lamina of sharks, reptiles, fishes and amphibians is active and constantly provides cells for the formation of new teeth. However, in humans a dental lamina is formed during early tooth development, but degenerates after the formation of the permanent teeth. If humans start with the same dental lamina as other vertebrates, why is this structure lost after permanent teeth formation? More importantly, to prevent the degeneration or recover the activity becomes a new strategy for biological tooth regeneration. Therefore, the lack of stem cells is just not the problem that prevents human teeth from regenerating. The loss of the dental lamina could be the missing link in tooth regeneration, instead. Thus, future regenerative dentistry should not only involve stem cell transplantation, but also maintenance or reactivation of the dental lamina. If the researchers can figure out how sharks and reptiles maintain their dental lamina throughout their lives, it may be possible to develop therapies that restore similar regenerative conditions in humans (Fraser et al., “Shark Tooth Regeneration”).

Evolutionary Trade-offs in Mammalian Tooth Regeneration

This raises an interesting evolutionary question : Why did mammals lose most of their ability to continuously regenerate their teeth, when many other vertebrates retained this ability? One is that mammals developed highly specialized teeth with precise occlusion, so they could efficiently chew a wide variety of foods. Good occlusion of upper and lower teeth is important for efficient mastication. The continual replacement of teeth seen in sharks or many reptiles would break this precise occlusion and feeding efficiency would be compromised. Thus, the evolutionary loss of tooth regeneration over the lifespan may have been a trade-off where mammals traded regenerative capacity for highly specialized and functional dentitions. Learning

how this evolutionary change happened could provide important clues for restoring regenerative capacity without compromising normal tooth function.

Tooth loss as a possible regenerative trigger

Another important question is whether tooth loss itself can activate tooth regeneration. Sharks grow replacement teeth all the time, whether they lose a tooth or not. This implies that the conveyor belt-like system in sharks is not driven by tooth loss. However, studies in alligators have demonstrated that the removal of teeth activates the dental lamina to initiate a new replacement cycle . This raises an important question: if tooth loss activates regenerative signaling in alligators, could the same type of signal be artificially activated in humans? If scientists can work out the molecular signals that initiate alligator tooth replacement, they might discover new methods for regenerating human teeth (Wu et al.)

The importance of timing and signaling control

Moreover, studies in vertebrate species have revealed that many of the molecular pathways governing tooth regeneration such as Wnt/β-catenin, BMP, SHH, FGF and Sox2 are evolutionarily highly conserved. But these pathways are not autonomous. Successful tooth regeneration relies on the spatial and temporal regulation of multiple signaling networks. For example, experimental activation of a single pathway such as Wnt signaling can even disrupt the typical pattern of tooth replacement, indicating that coordinated regulation is critical to maintain a functional regenerative system.

Limitations

Animal models provide important insights into tooth regeneration. However, there are several limitations. Differences in jaw structure, tooth morphology and the organization of the dental lamina prevent direct translation of findings from vertebrates to humans. While many developmental pathways are evolutionarily conserved (Wnt, BMP, SHH, FGF) their regulation is different from species to species and requires precise spatial and temporal control. In addition, most regenerative strategies have been tested in animal models or pre-clinical studies, and further studies are required to validate their long-term safety, functional integration and clinical efficacy in humans.

Future Perspective

Although remarkable progress has been made in stem cell biology, biomaterial engineering and molecular signaling, current regenerative strategies are still largely focused on regenerating individual dental tissues, rather than the biological system responsible for lifelong tooth replacement. Comparative studies of naturally regenerating vertebrates implicate the necessity of the specialized developmental environment of the dental lamina, in addition to the presence of stem cells, for successful tooth regeneration.

Therefore, future regenerative dentistry could move away from the transplantation of stem cells and focus on recreating or reactivating the developmental niche that inherently controls tooth formation. Emerging technologies such as artificial dental lamina, dental organoids, bioengineered epithelia and 3D-bioprinted stem cell niches may eventually provide an environment that is conducive to coordinated tooth development. The combination of these approaches with controlled signaling molecules and patient-specific stem cells may allow regeneration of whole functional teeth instead of individual dental tissues. Future regenerative dentistry may ultimately succeed by reactivating developmental programs conserved throughout vertebrate evolution but largely lost in humans, rather than trying to invent a whole new biological mechanism. Therefore, understanding the mechanism of how nature maintains lifelong tooth regeneration could be the basis for biological therapy for missing teeth replacement with fully functional natural teeth.

Despite important unresolved scientific and technical challenges, advances in developmental biology, biomaterials and tissue engineering indicate that biological tooth regeneration becomes an ever more realistic long-term goal.

Conclusion

Tooth regeneration is still one of the most significant challenges of regenerative dentistry, since teeth are highly organized organs with multiple specialized tissues that have to develop simultaneously. Despite scientific advances, it is still difficult to regenerate a complete functional human tooth. Comparative studies in vertebrates suggest that tooth regeneration is not as simple as one new tooth formation. It is the result of a whole biological system in which many parts function in a coordinated manner.

Current regenerative approaches have achieved regeneration of single dental tissues such as dentin, pulp, periodontal ligament and tooth-like structure. But future research should not regenerate isolated tissues, but rather reconstruct the biological system that normally regulates tooth development and replacement. In particular, information on the maintenance of the dental lamina and the initiation of the regenerative signal can provide clues for future biological tooth regeneration. In the end, the lessons learned from naturally regenerated vertebrates may not only enhance our understanding of tooth development, but also shed light on the development of safer and more effective regenerative therapies that can restore natural human teeth. Understanding how nature has evolved to maintain life-long tooth regeneration in vertebrates, may lead to future regenerative dentistry, which could allow humans to regenerate fully functional natural teeth rather than artificial replacements.

Works Cited

Chen, Haifeng, and Yihong Liu. “Teeth.” Advanced Ceramics for Dentistry, edited by James Zhijian Shen and Tomaž Kosmač, Butterworth-Heinemann, 2014, pp. 5–21. https://doi.org/10.1016/B978-0-12-394619-5.00002-X.

Davit-Béal, Tiphaine, et al. “Amphibian Teeth: Current Knowledge, Unanswered Questions, and Some Directions for Future Research.” Biological Reviews, vol. 82, no. 1, 2007, pp. 49–81. Wiley Online Library, https://doi.org/10.1111/j.1469-185X.2006.00003.x

Fraser, Gareth J., et al. “Shark Tooth Regeneration Reveals Common Stem Cell Characters in Both Human Rested Lamina and Ameloblastoma.” Scientific Reports, vol. 9, 2019, article 15956. https://doi.org/10.1038/s41598-019-52406-z.

Fraser, Gareth J., Ariane S. I. Standing, Christopher Underwood, and Alexandre P. Thiery. “The Dental Lamina: An Essential Structure for Perpetual Tooth Regeneration in Sharks.” Integrative and Comparative Biology, vol. 60, no. 3, 2020, pp. 644–655. https://doi.org/10.1093/icb/icaa102.

Jansen, J. A. “The Current State and Future of Oral Health: A Position Paper Exploring the Role of Regenerative Dentistry.” Regenerative Medicine and Dentistry, vol. 2, no. 1, 2025, article 6. https://doi.org/10.53941/rmd.2025.100006.

Juuri, Elina, et al. “Sox2+ Stem Cells Contribute to All Epithelial Lineages of the Tooth via Sfrp5+ Progenitors.” Developmental Cell, vol. 23, no. 2, 2012, pp. 317–328. https://doi.org/10.1016/j.devcel.2012.05.012.

Liu, Peng, et al. “Effects of SOX2 on Proliferation, Migration and Adhesion of Human Dental Pulp Stem Cells.” PLoS ONE, vol. 10, no. 10, 2015, e0141346. https://doi.org/10.1371/journal.pone.0141346.

Morita, K., et al. “The Next Generation of Regenerative Dentistry: From Tooth Development Biology to Periodontal Tissue, Dental Pulp, and Whole Tooth Reconstruction in the Clinical Setting.” Regenerative Therapy, vol. 28, 2025, pp. 333–344. https://doi.org/10.1016/j.reth.2025.01.002.

Murashima-Suginami, A., et al. “Anti-USAG-1 Therapy for Tooth Regeneration through Enhanced BMP Signaling.” Science Advances, vol. 7, no. 7, 2021, eabf1798. https://doi.org/10.1126/sciadv.abf1798.

Nakashima, Misako, and Takashi Tanaka. “Dental Pulp Regeneration Using Autologous Dental Pulp Stem Cells.” Regenerative Therapy, vol. 26, 2024, pp. 1–10.

Sui, Y., Zhou, Z., Zhang, S., et al. “The Comprehensive Progress of Tooth Regeneration from Tooth Development to Tissue Engineering and Clinical Application.” Cell Regeneration, vol. 14, 2025, article 33. https://doi.org/10.1186/s13619-025-00249-7.

Takahashi, K., et al. “Development of a New Antibody Drug to Treat Congenital Tooth Agenesis.” Nature Communications, 2024.

Thiery, Alexandre P., Kyle J. Martin, Katherine James, et al. “Shark Dental Lamina Molecular Regulation.” bioRxiv, 2025. https://doi.org/10.1101/2025.08.26.672487.

Tucker, Abigail S., and Gareth J. Fraser. “Evolution and Developmental Diversity of Tooth Regeneration.” Seminars in Cell & Developmental Biology, vols. 25–26, 2014, pp. 71–80. https://doi.org/10.1016/j.semcdb.2013.12.013.

Wu, Ping, et al. “Specialized Stem Cell Niche Enables Precursor Cell-Mediated Tooth Regeneration in the Alligator.” Proceedings of the National Academy of Sciences of the United States of America, vol. 110, no. 22, 2013, pp. E2009–E2018. https://doi.org/10.1073/pnas.1213202110.

Xuan, Ke, et al. “Deciduous Autologous Tooth Stem Cells Regenerate Dental Pulp after Implantation into Injured Teeth.” Science Translational Medicine, vol. 10, no. 455, 2018.

From Arrival to Integration: An Analysis of Political Trust and Civic Engagement Among Afghan, Ukrainian, and Venezuelan Refugees in the United States By Long Nguyen

Abstract

Scholarship on refugee integration has often viewed political trust as an outlook refugees bring with them from the political systems in their countries of origin, assuming that experiences under different types of governments continue to influence how refugees view public institutions after resettlement. However, this theory has rarely been examined across refugee populations from significantly different political backgrounds but living with the same institutions in one host country. Through a comparative analysis of Afghan, Ukrainian, and Venezuelan refugees arriving in the United States since 2015, this study finds that both political trust and civic participation are more influenced by the institutions migrants encounter upon arrival than the regimes from which they fled. Across all three case studies, changes in legal stability and sustained support from both government agencies and community institutions were most closely associated with increased naturalization and civic participation rates, whether in the electoral process or in established institutions. Although political experiences before migration somewhat influence refugees' initial expectations of how they will be treated in the U.S., these beliefs are constantly changed through repeated interactions with American governmental and civic institutions. Therefore, by placing political trust and participation at the center of civic integration, this paper argues that the long-term relationship between refugees and the American state is ultimately defined by the legitimacy and consistency of the institutions that receive and integrate.

Introduction

For over two centuries, the United States has been a haven for refugees fleeing war, famine, and widespread political unrest in their home countries. Millions of people worldwide have escaped declining democracies and brutal dictatorships in search of freedom. While the reasons for fleeing may vary, refugees often experience similar issues when resettling, including assimilation into their new society, uncertainty regarding their legal status, and efforts to establish trust in political institutions that might share little with the ones they originally fled. In response to increased conflict and instability abroad, the U.S. has admitted large numbers of refugees from Afghanistan, Ukraine, and Venezuela since 2015. All three nations represent different political paths that have caused mass displacement.

Afghanistan's US-backed government, established during the War on Terror in 2001, ended with the reinstatement of the Taliban regime in August 2021, as tens of thousands of Afghan troops and civilians fled the nation in a large-scale evacuation. Over 10 million Afghans remain displaced globally, with the majority living in neighboring Iran and Pakistan, while over 190,000 reside in the United States (UNHCR; Montalvo and Batalova). Russia's full-scale invasion in February 2022 displaced over 5.9 million Ukrainians and placed the country's emerging democracy under intense strain (USA for UNHCR). In the United States, the number

of Ukrainian arrivals has grown significantly through humanitarian parole, numbering over 221,000 in states such as Illinois, New York, California, and Washington by 2026 (Childress et al.; Folsom and Batalova). In Venezuela, over two decades of authoritarian repression under Hugo Chávez and Nicolás Maduro prompted one of the world's largest displacement crises, with nearly eight million Venezuelans fleeing the country since 2015. Venezuelan displacement has been mostly regional, with significant populations in Colombia, Peru, Ecuador, Chile, and Brazil (Amaya and Batalova). Several hundred thousand Venezuelans were granted protected status in the United States, until the status was recently terminated.

Each refugee group arriving in the United States is significantly affected by the type of government they were accustomed to, their direct experiences with both political and civic institutions, and the various resettlement programs and legal processes they face upon arrival. This study analyzes how these differences in origin, combined with the legal and resettlement pathways each group receives, affect their levels of political trust and civic participation in their new country.

Research Question and Literature Review

Modern research on refugee integration has focused heavily on cultural and economic contributions. Political integration, which refers to how refugees come to trust in or engage with the institutions of their host country, has received comparatively little attention. Throughout this paper, the term "institutions" refers to both formal government agencies responsible for refugee admission and legal status, such as the Department of Homeland Security, and local civic institutions that play a role in helping refugees adjust to public life, such as schools, libraries, and community organizations. In the areas studied, the existing body of scholarship has developed along two tracks.

The first concerns civic engagement, driven mainly by the legal conditions and organizational methods present when a group arrives in the U.S. The Civic Voluntarism Model, developed by political scientists Sidney Verba, Kay Lehman Schlozman, and Henry E. Brady, provided a structured framework to explain some of the core reasons behind volunteering and overall participation in American politics. The Model found that involvement in community organizations creates habits of increased confidence and coordination that translate to greater engagement in voting and political advocacy overall (Verba et al. 4). In a comparison of immigrant incorporation in the U.S. and Canada, formal acquisition of citizenship and government-backed sponsorship were found to have measurable increases in civic participation, relative to communities living with precarious legal statuses or little government support (Bloemraad 677). Among undocumented immigrants in California, legal status alone did not completely explain lower rates of civic participation, but the political climate of an individual's county of residence also had a significant impact on these trends. For instance, immigrants with uncertain statuses living in counties with stronger Republican support in 2016 were overall less likely to participate in community organizations, a choice likely attributable to migrants' responses to an adverse political situation (Beck and Shklyan 12).

The second track involves political trust or confidence in the host country's institutions. It has been held in this literature that migrants who lived under authoritarian rule generally carry a lower sense of confidence in government into their new country, as their previous experiences amplified their perception of these institutions as lacking accountability or serving as a rubber stamp for repressive measures. A review of the literature found that the evidence was considerably more nuanced. Some studies of migrants from authoritarian backgrounds found that they held higher institutional trust in their host countries than even native-born citizens, as they were comparing a government that actively worked to repress them instead of one that simply fell short of delivering on its obligations to the people (Godinho Nascimiento et al. 15, 17). An earlier, broader review of the overall literature on political trust by Margaret Levi and Laura Stoker, prominent scholars of political institutions in democratic nations, establishes a clear contrast between general trust in the government and confidence in a particular institution, such as a government agency or other bureaucratic body (491). This distinction has not been applied specifically to refugee populations, although they are one of the groups most likely to have contact with their host government through a singular agency that temporarily handles resettlement and protection issues. Current research on various diaspora communities has, to some extent, documented how the capacity for civic organization develops within a population over time. Research on participants in Ukraine's 2004 Orange Revolution and 2013-14 Euromaidan protest movements found a civic culture that grew much stronger and more confident across the two movements, as both succeeded in bringing about significant change to the government, and many Ukrainians felt a strong obligation to their country (Thomson 9). Euromaidan, in particular, initiated an unprecedented wave of solidarity among the public, as spontaneous, informal community networks were quickly formed to aid protesters and advocate against many problems, such as corruption, that had long plagued Ukrainian politics. In the aftermath of the protests, activists mobilized to push through reforms and legislative priorities with the new government in place. (Shapovalova and Burlyuk 11). A more recent study discusses how this previous civic organization affected the Ukrainian diaspora's response to Russia's 2022 invasion through the Ukrainian World Congress, which engaged in an intensive international advocacy campaign through both physical and digital means immediately after the invasion began (Kozachenko 82).

However, none of the aforementioned studies directly compare refugees who represent considerably different political trajectories but arrive in a single host country within the same period. Afghanistan, Ukraine, and Venezuela are states with useful points of comparison, as each represents a different path that spurred some of the world's most significant refugee crises: the collapse of a government after years of outside intervention, the complete disruption of a nation by foreign invasion despite functioning institutions, and the continued decline of democracy through a decades-long authoritarian consolidation of power.

Framework and Methodology

This study combines legal analysis with theories of political trust and civic participation to examine how refugees view democratic institutions after resettling in their host country. It tests whether political trust is better explained through the interactions between refugees' previous political experience and their later encounters with institutions in the United States. The analysis uses civic voluntarism theory, research on immigrant political participation, the institutional and reference point theories of political trust, and legal scholarship on refugee protection to establish a method for evaluating how experiences with institutions affect civic integration in the long term. A comparative case study approach was employed, using Afghan, Ukrainian, and Venezuelan refugees as migrants from three nations with distinct political backgrounds and conditions that caused their respective refugee crises. It combines qualitative and quantitative evidence, including records from Congress, federal immigration statistics, public opinion surveys, comparative international research, diaspora records, and one interview with an Afghan refugee, to compare how legal pathways and experiences with institutions affect political trust and civic participation. Patterns of political trust were evaluated using observable metrics, including naturalization rates, political participation patterns, public opinion, civic organizations, and refugees' descriptions of their interactions and views of both governmental and community institutions in the United States. This design allows the paper to compare how political trust and civic engagement shift over time in refugee populations that faced different political circumstances prior to their migration.

Paper Map

This paper is organized into four sections, followed by a conclusion section. Section I discusses the initial political conditions that led each population to flee its country of origin, comparing the rapid 2021 collapse of Afghanistan's US-backed republic, Russia's 2022 invasion of Ukraine, and Venezuela's gradual authoritarian consolidation of power. Section II analyzes the legal pathways each population encountered after arriving in the United States, including the scope of different parole and humanitarian programs and the duration in which they were active. Section III compares the patterns of naturalization, civic participation, and political engagement through electoral trends and the development of organizations across all three refugee groups. Section IV applies the systems of political trust discussed earlier to examine the integration policies previously implemented by Canada, Colombia, and the U.S., and the correlations between these policies and refugees' institutional views. The conclusion then assesses whether the best explanation for the patterns of political trust and civic engagement identified throughout the paper involves the conditions in the nation of origin, the legal pathways refugees encountered upon their arrival, or an interaction between the two factors. It also identifies various unresolved questions that may represent directions for future research to inform refugee policy and legislation in the United States.

Section I: Political Conditions Leading to Mass Displacement

Political socialization theory holds that, although familial relationships are the strongest

influence on one's political beliefs, early encounters with a government's institutions also significantly affect expectations of how actively citizens believe they should participate in their government, such as their civic duties and their views of the government's responsibilities to the people in return (Almond and Verba). Afghanistan, Ukraine, and Venezuela are three distinct types of these encounters.

Amidst the Bush administration's War on Terror from 2001 onwards, the Islamic Republic of Afghanistan was installed with direct military and financial aid from the United States and its allies. A constitution was introduced in 2004, followed by several rounds of democratic elections. However, even in its early years, the central government lacked the capacity to extend its authority beyond major cities and struggled to control the influence of powerful warlords, as it relied heavily on these strongmen to combat the Taliban insurgency in remote regions (Rubin 9). Furthermore, widespread corruption within the military and bureaucracy severely weakened the government's legitimacy over its two decades in power. In 2019, the Trump administration entered into direct negotiations with the Taliban, resulting in the 2020 Doha Agreement and the eventual withdrawal of NATO forces from Afghanistan. As the withdrawal got underway, the Taliban quickly launched a major offensive in the spring of 2021 and captured the capital, Kabul, by August. By that point, the Afghan National Army had largely dissolved, and President Ashraf Ghani fled to the United Arab Emirates, sparking the largest evacuation in U.S. military history.

Ukraine's political development took a considerably different course but began in a similar period. The Orange Revolution, triggered by accusations of electoral fraud in the 2004 Ukrainian presidential election runoff, was characterized by sustained street protests and widespread acts of civil disobedience, eventually leading to a re-vote of the runoff. A decade later, the Euromaidan protest movement (2013-2014) saw a similar pattern after the government of Viktor Yanukovych abruptly ended negotiations on a European Union association agreement and opted for closer ties to Russia, once again driving large-scale change and forcing the resignation of the government. In February 2014, Russian forces invaded and occupied Crimea, a peninsula located south of Ukraine. Soon after, Russian-backed paramilitaries staged uprisings in Ukraine's eastern Donbas region and subsequently declared the Donetsk and Luhansk People's Republics to be independent states. International agreements such as the Minsk Protocol and the Normandy format attempted to establish a ceasefire and resolve the conflict, but disagreements between Russia and Ukraine regarding elections and other issues stalled the peace process significantly (Walker 11-16). Throughout 2021, Russia aggressively pursued a buildup of military forces along its border with Ukraine, and launched a full-scale war in February 2022, abruptly interrupting a democracy that had undergone two mass civic uprisings in just the past decade.

Venezuela's emigration, one of the largest globally, arose from the consolidation of power and subsequent economic crisis under two successive authoritarian governments. The nation underwent institutional change under the leadership of Hugo Chávez, beginning with his election in 1998 and the 1999 amending of the Constitution, which concentrated authority in the hands of

the executive. Despite initially enjoying popular support, Chávez's regime employed heavy-handed tactics and security forces to consolidate power, particularly through the subversion of the judiciary and the suppression of both the media and political opponents. For instance, the Supreme Tribunal of Justice, the highest court in Venezuela, was restructured by adding justices who were aligned with the government. Long-standing independent media outlets such as *Radio Caracas Televisión (RCTV)* lost their broadcast licenses, and several national elections conducted under Chávez's rule were criticized by international observers as marred by fraud (Corrales and Penfold). Until his capture by the United States military in January 2026, the government of Chávez's successor, Nicolás Maduro, presided over widespread repression, human rights abuses such as arbitrary detention, torture, and forced disappearances, as well as a currency collapse and hyperinflation reaching 1,300,000% by 2018 (Hanna; Waldron 1). The regime's continued erosion of democratic values and institutions through electoral manipulation and media control essentially destroyed its legitimacy, leading to widespread distrust among the population. By 2019, these conditions had driven nearly 8 million people from the country (Freier and Parent 61).

Section II: Legal Pathways and Resettlement Programs

The legal pathways each group of refugees received are the main intermediaries between their prior political experiences and the government that now hosts them. All represent different versions of liminal legality, the condition of having an ambiguous, "in-between" immigration status. These types of legal statuses leave many refugees in a state of ongoing uncertainty; instead of providing the lasting security that green cards or citizenship would, they require repeated renewal, making long-term decisions regarding employment and housing more difficult (Menjívar 1001).

Afghan refugees entered the United States through several legal pathways, although the largest group arrived through humanitarian parole or Special Immigrant Visas (SIVs) under the U.S. government's evacuation and resettlement efforts. Upon the fall of Kabul in 2021, Operation Allies Welcome brought about 97,000 refugees to the U.S., with 77,111 receiving parole. Following this initial surge, over 180,000 Afghans were admitted and resettled across the nation by 2025 (Office of Inspector General; Weinbaum and Ali). These programs granted parole for two years at ports of entry and did not include a pathway to obtaining a green card or citizenship in the U.S. Recognizing this, Congress attached a provision to a 2021 appropriations act requiring asylum applications filed by Afghan refugees to be decided within 150 days (Parness). However, the average processing time for these claims has reached over 300 days. By late 2023, federal courts had ordered U.S. Citizenship and Immigration Services (USCIS) to resolve part of the backlogged cases that remained in extended security screening. Congress has repeatedly attempted to pass legislation to streamline the path from parole to permanent residency. For instance, the Afghan Adjustment Act would grant lawful permanent residency to Afghan parolees who arrived in the U.S. as evacuees after a process of enhanced vetting, including biometric data collection and interviews. Furthermore, it would widen the eligibility

for Special Immigrant Visas to Afghan soldiers who fought alongside American forces, including members of the Afghan Air Force and Female Tactical Teams (Haidari). No version of the bill has passed either chamber of Congress; therefore, continually renewed parole is the only legal pathway available to most Afghans in the United States.

Ukrainian refugees were allowed entry through Uniting for Ukraine, a humanitarian program launched in 2022 that required a U.S.-based sponsor to file a Form I-34 Affidavit of Support for up to two years before parole was granted (Federal Register). Nearly 240,000 Ukrainians were paroled through Uniting for Ukraine alone, with an additional 30,000-35,000 directly granted parole by border officials (Ukraine Immigration Task Force). An analysis of the zip codes of sponsors found that congressional districts with the most sponsorship applications had a Ukrainian ancestry rate 219% greater than the national average. Sponsors also tended to lean Democratic by a large margin: 68% of sponsors resided in districts represented by House Democrats compared to just 30% in Republican districts. Several Republican districts, however, such as those of Representatives Brian Fitzpatrick of Pennsylvania, Nicole Malliotakis of New York, and Jaime Herrera Beutler of Washington, were among the areas with the most active sponsorship in the nation (Niskanen Center). Until Congress passed legislation making Uniting for Ukraine parolees eligible for federal resettlement funding and other forms of government assistance, they were generally excluded from these benefits, but were authorized to begin working immediately using I-94 documentation while awaiting their Employment Authorization Documents (U.S. Citizenship and Immigration Services, "Certain Afghan"). At the start of the second Trump administration in January 2025, the program was suspended, and both new and pending sponsorship applications were completely halted. A four-year retrospective published in 2026 describes many parolees as facing a variety of problems that had surfaced since the suspension, including backlogs in the processing of work permit renewal applications, rising fees for filing documents, and a near-total stop in the approval of Temporary Protected Status (TPS) applications for Ukrainians in the United States. The report also noted that most pathways to a long-term status, such as green cards or temporary work visas, were practically unavailable to Ukrainian parolees, with some requiring an in-person interview at the U.S. consulate inside Ukraine itself (Ukraine Immigration Task Force).

Venezuelan refugees relied largely on the aforementioned Temporary Protected Status, a procedure established by Congress as part of the Immigration Act of 1990 and as a replacement for the prior Extended Voluntary Departure system, seen by many lawmakers as lacking guardrails and a clear structure (Leblang et al.). TPS allows for legal protection in eighteen-month intervals, renewed at the discretion of the Secretary of Homeland Security. The 1980 Refugee Act restricted protection to individuals matching the formal refugee definition of the United Nations, but left no clear status for those fleeing conditions such as civil war or natural disasters, for whom TPS sought to provide a pathway of safe haven (Frelick 43). The Department of Homeland Security (DHS) can designate a foreign country for TPS if those conditions make a migrant's return to their nation unsafe, and allows for work authorizations and protections from deportation for a certain period, but does not provide a path to permanent

residency. In March 2021, Venezuela was first designated for TPS; this designation was extended and redesignated in September 2023, when DHS reported approximately 242,700 beneficiaries of the program and identified 472,000 more Venezuelan nationals as eligible to apply (U.S. Citizenship and Immigration Services, "DHS Previews"). By January 2025, the total number of active Venezuelan TPS beneficiaries had reached 607,000, and the outgoing Biden administration had extended the 2023 designation through October 2026 before leaving office (U.S. Citizenship and Immigration Services, "DHS Publishes"). A separate humanitarian parole program, the Processes for Cubans, Haitians, Nicaraguans, and Venezuelans (CHNV), admitted an estimated 117,000 additional Venezuelans from 2022 onwards (U.S. Customs and Border Protection). In October 2025, the incoming Trump administration moved to reverse the January extension and terminate TPS for Venezuela, citing significant national security threats from Venezuelan gangs such as Tren de Aragua. Within about five months, 350,000 Venezuelans covered by the 2023 designation were notified of multiple successive changes to their legal status. First, they were told that their designation was essentially secure through 2026, informed in early February that it would expire in approximately two months, and finally notified in May that the Supreme Court had allowed the termination to proceed amid litigation (U.S. Citizenship and Immigration Services, "Update: Supreme Court Order"). The earlier 2021 TPS designation for Venezuela expired and was subsequently terminated by DHS in November 2025.

Section III: Trends in Community and Civic Engagement

Civic participation of Afghan refugees in the United States commonly operates through community-based organizations rather than more formal methods of political advocacy. Groups such as the Afghan American Foundation and the Afghan Coalition often provide cultural programs and legal referrals for many migrants (Center of Excellence for Immigrant Child Health and Wellbeing). For instance, in 2025, the Afghan-American Community Organization's annual conference drew over 400 attendees, distributed $30,000 in scholarships, and conducted programs on legal education reaching over 2,220 participants (Afghan-American Community Organization). As many Afghan refugees have experienced conditions such as depression and PTSD, community organizations and support networks have been pivotal in improving their mental health by connecting them with the broader Afghan diaspora and helping them learn English to gain stable employment. Still, many resettlement agencies lack sufficient government funding and training needed to address refugees' needs, particularly those of Afghan women (Aghaei et al. 82). Beyond participation in community organizations, naturalization is another significant indicator of Afghan refugees' involvement in American civic life. Seventy-two percent of the current Afghan immigrant population arrived in the U.S. after 2010, and the group's naturalization rate is about 37 percent, compared to 53 percent for the overall foreign-born population (Montalvo and Batalova). An administrative study of refugees resettled in the U.S. between 2000 and 2010 by Nadwa Mossaad et al., migration policy researchers who examined refugee integration in the United States, found that they typically naturalized more quickly than other lawful permanent residents, with 66 percent receiving American citizenship

by 2015. It identified factors such as the length of residency, educational level upon arrival, and settlement in urban areas with larger populations of the same ethnicity as strong signs of naturalization (Mossaad et al. 9176). As most Afghans currently in the U.S. arrived through parole programs, and the largest group arrived only in 2021, it is likely that the relatively low naturalization rate is attributable to a population still maneuvering the citizenship process. Over time, civic participation would be expected to increase as well.

Engagement among the Ukrainian diaspora predates Russia's 2022 invasion by more than eighty years. A national federation founded in 1940, the Ukrainian Congress Committee of America, brought together approximately thirty different member organizations and currently represents an estimated one million Americans of Ukrainian descent. Established in September 2022, the American Coalition for Ukraine also eventually expanded to include over one hundred U.S.-based member organizations (Sõukand 9, 10). In addition to these umbrella organizations, existing nonprofit groups intensified their efforts following the initial influx of Ukrainian refugees in March 2022. Razom for Ukraine, a nonprofit organization founded to provide humanitarian assistance and support for Ukrainians, raised over $100 million from 120,000 individual donors and maintains a permanent advocacy office in Washington, alongside both the Congressional Ukraine Caucus and a similar Senate Caucus (Popov). Organizational networks also engage in more direct political advocacy. During the 2023-2024 election cycle, political action committees (PACs) aligned with Ukrainian interests such as The American Ukraine PAC contributed roughly $40,100 to congressional candidates, with about three-quarters of that amount going to Republican candidates (Popov). Formal lobbying, however, is just one aspect of civic participation among Ukrainian refugees, as much of this political activity occurs outside the electoral process and does not depend on U.S. citizenship. For instance, rallies supporting Ukraine have been held near the Capitol continuously since the start of the invasion. Anniversary demonstrations held at the Lincoln Memorial and other venues, organized by the Embassy of Ukraine in conjunction with other diaspora organizations such as Razom for Ukraine and United Help Ukraine, have drawn thousands of participants since February 2023 (Jacome). These activities point to a highly visible and organized public presence, yet relatively little scholarship has examined how Ukrainian refugees themselves participate in civic life within the U.S., beyond documenting the organizations that support them. Moving past just receiving assistance, direct participation in advocacy efforts and local decision-making was closely associated with stronger resilience among displaced Ukrainians, as establishing and maintaining critical social ties through community associations helped reduce the effects of previous trauma (Migliorini et al. 2). Among refugees resettled from 2000 to 2010, individuals from Ukraine, along with those from Burma, Vietnam, and Liberia, naturalized at a rate 11 to 21 percent lower than the overall refugee population (Mossaad et al. 9177). Although this group of migrants arrived years before the current conflict and may not be representative of today's refugees, the findings still show that extensive civic engagement efforts can take place even with lower naturalization rates. A study of the Ukrainian diaspora in France and Poland similarly found that, despite disparities in the size and development of existing community networks, the Russian invasion galvanized

Ukrainian advocacy efforts and produced extensive mobilization in both nations. Many Ukrainian activists quickly organized demonstrations and other public campaigns, but the crisis also notably drew less civically engaged Ukrainians living in France into community and political activities to support their home country (Fihel et al. 706, 716).

Data on Venezuelan refugees show that active participation in elections and a decline in the popularity of certain political leaders have both occurred over the past few years. Between 2008 and 2018, the eligible voting population of Venezuelan Americans in Florida grew by 184%, increasing from about 26,000 to over 75,000, yet accounting for just about 2% of the state's Hispanic electorate in 2020 (Noe-Bustamante). To put the figure into context, about 47% of all Venezuelan immigrants in the U.S. live in Florida, and the group has a significantly higher rate of college education than the overall foreign-born population (Amaya and Batalova), yet another factor identified by Mossaad et. al. as a reliable indicator of naturalization. With regards to elections, in Miami-Dade County, the vote share of the Republican presidential candidate increased from 33.8% in 2016 to 55.2% in 2024, while that of the Democratic presidential candidate declined from 63.2% to 43.71 percent over the same period (NBC6). However, this political realignment proved to be much briefer than the election results suggested, as substantial dissatisfaction with elected officials soon became evident. In a 2025 survey of Venezuelans in Florida, 80% of respondents, including half of those who had become citizens, viewed the termination of TPS as unjust. The majority of respondents who voted for Donald Trump in 2024 expressed either regret or mixed feelings about their decision within a year (Padgett). Similarly, 60% of respondents viewed Secretary of State Marco Rubio unfavorably, and 41% of eligible voters indicated that they would not support their current congressional representative in the 2026 midterm elections (Padgett).

Section IV: Political Trust and Policy Implications

Political trust refers to an individual's confidence that government institutions and officials in a nation will act competently and follow through on the commitments they make in the public interest (Turper and Aarts 417; Levi and Stoker 485). As previously mentioned, trust in any particular institution is distinct from trust in the government. Using survey data from ten countries in the post-communist era, political scientists William Mishler and Richard Rose, whose work analyzes how institutions affect public attitudes towards government, compared two competing explanations for political trust. Cultural theories argue that trust is determined by a person's political values and their experiences with government early in life, while institutional theories hold that trust is dependent upon the performance of political institutions as time goes on. Their results generally favored the latter explanation, as trust was more closely linked to recent government performance than to political conditions under previous regimes (Mishler and Rose 31, 37, 55). Personal interactions with these institutions were also stronger indicators of political trust than the nation's overall political climate. Furthermore, Johannes Röder and Jan Mühlau, sociologists who study the political experiences of immigrant populations, reached a similar conclusion in their examination of political trust among migrants. Their analysis used

pooled data gathered in Round 4 of the European Social Survey, which was conducted from 2008 to 2009. Immigrants reported higher levels of trust in public institutions than native-born citizens, a pattern attributable to both a "reference point effect" and the effectiveness of the host country's governance. Typically, immigrants compare the institutions in their host country against those in their home country; therefore, institutions that seem ordinary or even flawed to native-born citizens may be a significant improvement for individuals coming from authoritarian or corrupt political systems. The reference point effect weakened over time and across generations, as first-generation immigrants exhibited higher trust in institutions, whereas second-generation immigrants gradually became more skeptical (Röder and Mühlau 779, 780, 791, 792). As migrants gained more direct experience with a nation's institutions, their patterns of trust were mostly aligned with the institutional theory. Taken together, these systems strongly suggest that levels of political trust among refugee groups depend on two main factors: the institutions immigrants compare the United States against and their own, individual experiences with the American government after their arrival. The evidence in this paper cannot be explained by the country of origin alone; instead, political trust shifts as migrants' initial assessments are either replaced or strengthened by those experiences.

Prior to 2021, as Afghanistan's former US-backed government was democratically elected and formally operated under a Constitution, Afghan refugees who lived under the Islamic Republic would be expected to have a different point of comparison and weaker reference point than those from a long-standing authoritarian regime. However, the evidence points elsewhere. Afghan naturalization rates still remain far below the historical average for refugees, attributable to years of consistent delays in the United States' asylum system. Similar to Afghanistan, Ukraine emerged as a developing democracy, even with shortcomings in both nations such as widespread corruption.

Following the 2014 Revolution of Dignity, Ukraine held relatively free parliamentary elections and successive administrations have made efforts to implement reforms and improve democratic institutions (Krawchenko). Therefore, as Ukrainian refugees compared institutions in the United States with those of a duly elected government, the reference point effect should have been very limited, making direct experience with American institutions more important. Ukrainian refugees had less reason to view American institutions as a significant improvement over those they left behind. Furthermore, as programs such as Uniting for Ukraine were a stable source of government aid for refugees, the recent termination of the program suggests that there will be a considerable impact on their political trust.

Venezuela provides perhaps the clearest example of how the reference point effect can drastically change over time. After two decades of democratic decline, governmental institutions in Venezuela became little more than instruments of repression and power consolidation. Compared to those experiences, American institutions would initially seem much more trustworthy and thus produce a strong reference point effect. However, the evidence in Section III indicates that these views faded as Venezuelan refugees increasingly accumulated their own experiences with American institutions, most prominently the deportation policies of the Trump

administration. Wide shifts in electoral support were soon followed by a decline in political trust in about a year, changes that occurred too rapidly to be characteristic of established political attitudes carried over from Venezuela. These trends are consistent with both the institutional theory analyzed by Mishler and Rose, as well as the reference point hypothesis by Röder and Mühlau. Other nations have previously faced similar questions about how to respond to large influxes of refugee populations after major conflicts or humanitarian disasters. Canada and Colombia each used different legal approaches during a similar time period examined in this paper, and both have been the subject of a substantial body of research. If political trust is developed mainly through direct experience with governmental institutions, then the way governments design and oversee refugee programs is particularly important. Canada's Private Sponsorship of Refugees Program, which has been operating since 1979, has resettled over 325,000 refugees through sponsorship arrangements (Hyndman et al. 1). Unlike Uniting for Ukraine, without government assistance, a group of private citizens would take on the financial and housing responsibilities for a newly arrived family, usually for a one-year period. A ten-year longitudinal study of Southeast Asian refugees who arrived in Canada between 1979 and 1981 found that migrants resettled through sponsorship programs saw greater success in integration than those who received government assistance alone (Beiser). Specifically, privately sponsored Vietnamese refugees in Canada also had higher rates of employment and income than refugees assisted by the government for up to 15 years after their arrival, even after accounting for measurable differences between the two groups. This success was also supported by a rapid acquisition of language; through Canadian education programs and tutoring, the majority of Vietnamese refugees gained a basic understanding of English within 1-2 years after arrival (Hou 4819, 4823).

These results were corroborated by a subsequent study on Syrian refugees resettled across six Canadian cities, in which the type of sponsorship and the presence of a strong network of acquaintances had measurable impacts on the extent of civic integration, after controlling for differences between families (Hynie et al. 46-47). Much of that difference was attributed by the researchers to the longstanding relationships between refugees and their private sponsors. As refugees adjusted to life in Canada, their sponsors continued to provide guidance and daily support, unlike government assistance, which typically ends after a certain period of resettlement. In this instance, sustained contact was representative of the repeated interactions between institutions and communities that helped to build a sense of trust. Uniting for Ukraine adopted the idea of sponsorship but did not include the features most critical for increasing trust among migrants in Canada. For instance, Canadian sponsors committed to supporting refugees for the one-year period, but they arrived as lawful permanent residents instead of parolees under a temporary status. Uniting for Ukraine, however, required sponsors to submit a financial affidavit for refugees granted two years of parole without a pathway to permanent residency. A smaller American program launched in October 2021, The Sponsor Circle Program for Afghans (SCPA), had previously attempted to bring a version of the Canadian model into the United States by organizing groups of at least five American sponsors to support a family's resettlement,

successfully relocating approximately 600 Afghans. Although they provided personal and community benefits, Sponsor Circles had significant limitations, such as a 90-day minimum support time, a period much shorter than that of federal resettlement agencies. Furthermore, as most Afghan refugees arrived under parole, their rights were limited from the outset, and the program simply could not resolve the underlying issue of obtaining citizenship (Prantl 60-1).

Colombia's response to Venezuelan displacement provides one of the most direct comparisons to the TPS program in the United States, as both nations faced a similar crisis during the same time period. In February 2021, the Colombian government, under the Duque government, created a ten-year Temporary Protection Status, known as the *Estatuto Temporal de Protección*, for over 1.7 million Venezuelan migrants, most of whom had previously lacked any legal status in the country (Rojas et al. 1, 11). Using administrative employment records, economists observed that the roughly 281,000 Venezuelans who received legal status through Colombia's previous *Permiso Especial de Permanencia* (PEP) regularization program entered the workforce at much higher rates than those without permits. The ten-year protection program did not have a significant impact on wages earned, hours worked, or participation in the labor force among native Colombian workers while benefiting Venezuelan migrants (Bahar et al. 4). Moreover, Ibáñez et al., researchers and economists who studied forced migration and internally displaced persons (IDPs), found that the PEP program actually led to marked improvements in migrants' overall economic prosperity and health ("Life Out of the Shadows" 944). Subsequent research showed an increase in reported crimes against migrant women following the regularization process. The prevalence of crimes did not necessarily increase, but the women's new legal statuses increased their confidence in law enforcement and their ability to seek protection, therefore likely boosting their political trust as well. Simultaneously, overall domestic crime declined as regularization gave victims safer ways to report abuse (Ibáñez et al., "Protecting the vulnerable" 2-3).

The main distinction between the American and Colombian governments' responses to the crisis was the length of protection for migrants. Colombia's program was created by presidential decree but guaranteed ten years of legal protection from its inception, thus giving migrants an extended period of time to find employment and plan for their futures. TPS, however, offered only temporary relief dependent upon frequent renewals, so legal protections were often vulnerable to policy changes from one administration to the next. The evidence examined suggests that governments can increase political trust and civic engagement among refugees by improving the quality of the public institutions refugees encounter upon their arrival in the host country. If trust is increasingly established through repeated interactions with institutions, policies should seek to make them more transparent and consistent. Governments should first strive to reduce uncertainty about refugees' legal statuses over an extended period of time when possible, as ambiguity can negatively affect migrants' economic prosperity and general success in their host country. A study on asylum applicants in Switzerland found that every additional year an applicant waited for their asylum decision reduced their subsequent employment rate by four to five percentage points, a decline of about 16 to 23 percent compared

to the average employment rate in the nation (Hainmueller et al. 1-2). Although the study was limited to employment, it demonstrated how the psychological toll on migrants, originating from prolonged uncertainty about legal status, affected their job prospects. Greater emphasis should also be placed on transparency and fairness throughout the immigration process, as research on procedural justice indicates that people are more likely to comply with government decisions if they believe the rulings were made impartially, even if the outcome is unfavorable. This principle can likely also be applied to migrants, who may be more willing to accept and cooperate with legal decisions in general if they view the processes behind them as legitimate (Tyler 309, 312).

Furthermore, local institutions where refugees are most likely to interact with the public should also receive more support. Schools, libraries, neighborhood organizations, and volunteer programs, for instance, provide refugees with opportunities to meet residents and familiarize themselves with daily life. A 2019 study drawing on 22 in-depth interviews conducted with people involved in German nonprofit organizations serving refugees found that daily interactions between immigrants, refugees, and native-born citizens increased social trust and allowed for stronger community integration, especially because all groups were willing to adapt to each other's cultures for the common purpose of working to benefit aid recipients (Ruiz Sportsmann and Greenspan 938). The importance of institutions also emerged during an interview conducted for this paper with an Afghan refugee who arrived after the 2021 evacuation. Describing the loss of resettlement support from the federal government, he explained that "all the organizations stopped," leaving a local YMCA to cover "two months' rent of my apartment" before that assistance also ended (Interviewee 1). Although support organizations cannot replace government policies, they can establish the foundational relationships necessary for refugees to develop trust in public institutions.

Another approach focuses on expanding opportunities for local civic participation. A systematic review combining findings from 92 peer-reviewed articles and institutional reports published between 2000 and 2024 found that participatory budgeting, a process adopted by over 7,000 municipalities worldwide that allows residents to aid in decisions regarding how portions of a municipal budget are spent, has also been associated with increasing levels of political trust by making the government more accountable and responsive to a community's concerns (Islam 117, 118). Participatory budgeting has already been implemented in various communities throughout the U.S. to help immigrant populations better integrate into public life. For instance, in New York City, the process effectively served as an entryway to civic engagement for Bangladeshi migrants, especially when issues they were passionate about gave them a direct reason to advocate within their communities and make their ideas known (Hayduk et al. 91, 92).

Conclusion and Findings

This paper examined how political trust and civic participation among Afghan, Ukrainian, and Venezuelan refugees are affected by both the political systems they fled from and the legal pathways they faced after their arrival in the United States. The evidence indicates that neither factor alone properly explains the patterns observed; political trust is instead developed

through the interactions between migrants' previous political encounters and their direct experiences with the institutions of their host country. The recent raid that deposed Venezuelan President Nicolás Maduro demonstrates how these comparisons continue to change even after migration, and is a significant development for Venezuelans both inside and outside the country. Most Venezuelans, unlike Ukrainians or Afghans, experienced longstanding internal repression without a complete transition to armed conflict. However, despite Maduro's removal from power, his regime remains deeply entrenched in the country and has faced continuous diplomatic and military pressure from the Trump administration. This has caused widespread speculation that Venezuelan migrants may be encouraged to return to their home country if the interim government relaxes its crackdown on dissent, perhaps also leading to changes in Venezuelans' trust in civic institutions to protect them. While such political conditions in a refugee's country of origin do affect the expectations refugees initially carry with them, these beliefs do not remain fixed. In the United States, for instance, refugee policy has been approached as a series of individual legal decisions that change often depending on the administration in office, requiring many refugees to adjust to an uncertain future. As time went on across all three case studies, even refugees who had more favorable views of American institutions saw their confidence reduced when faced with extended, uncertain legal processes and a lack of community or state support. In contrast, legal stability and consistent, visible efforts at transparency by governmental institutions were most closely associated with increases in political trust. These findings indicate that, although governments cannot alter or determine the political experiences faced by migrants in the past, they can take steps to improve the institutions refugees encounter upon their arrival and help strengthen their confidence in a host country's governance. The three refugee populations examined here suggest that refugees primarily assess the legitimacy of institutions and their overall relationship with the government through legal pathways. Therefore, political trust is both an outcome of successful civic integration in a state and one of the primary forces driving it.

Several inquiries remain open, and future research could examine if the patterns observed last beyond one or two generations of immigrants, or whether the American-educated children of refugees develop similar attitudes regarding civic engagement as their parents. In these instances, longitudinal studies would be useful in determining the changes in institutional trust, particularly if legal pathways are stabilized or if political conditions change rapidly in either the host country or the country of origin. It is critical to note that there are still an abundance of ongoing conflicts that have resulted in unprecedented levels of global displacement, such as civil wars in Sudan and Myanmar. Governments at every level will continue to confront the challenges of integrating individuals whose sociopolitical experiences may be drastically different from those of native-born citizens. Therefore, it is important to improve key institutions that can affect a government's legitimacy. A thorough understanding of how political trust is developed and how civic engagement is sustained is fundamental to strengthening the cohesion of American society and ensuring that its democratic institutions, viewed by millions as the free world's beacon of hope, continue to command the confidence of the public across an increasingly diverse nation.

Works Cited

Aghaei, A., et al. "Navigating Organizational Challenges and Formal Support for Afghan Refugee Women in California Amid Covid-19 and the Taliban's Takeover." International Journal of Social Determinants of Health and Health Services, vol. 56, no. 1, July 2025, pp. 81–95, https://doi.org/10.1177/27551938251360646.

Almond, Gabriel A., and Sidney Verba. "The Civic Culture: Political Attitudes and Democracy in Five Nations." Jstor, Princeton University Press, 1963, https://www.jstor.org/stable/j.ctt183pnr2.

Amaya, Ana Alanis, and Jeanne Batalova. "Venezuelan Immigrants in the United States." Migration Policy Institute, 6 Feb. 2025, https://www.migrationpolicy.org/journal/spotlight/venezuelan-immigrants-united-states. Accessed 14 July 2026.

Bahar, Dany, et al. "Give Me Your Tired and Your Poor: Impact of a Large-Scale Amnesty Program for Undocumented Refugees." Journal of Development Economics, vol. 151, Mar. 2021, p. 102652, https://doi.org/10.1016/j.jdeveco.2021.102652.

Beck, Kevin, and Karina Shklyan. "Civic Engagement, Legal Status, and the Context of Reception: Participation in Voluntary Associations among Undocumented Immigrants in California." Socius: Sociological Research for a Dynamic World, vol. 7, Jan. 2021, p. 237802312110052, https://doi.org/10.1177/23780231211005214.

Beiser, Morton. "Sponsorship and Resettlement Success." Journal of International Migration and Integration / Revue de l'integration et de La Migration Internationale, vol. 4, no. 2, Dec. 2003, pp. 203–215, https://doi.org/10.1007/s12134-003-1033-z.

Bloemraad, I. "Becoming a Citizen in the United States and Canada: Structured Mobilization and Immigrant Political Incorporation." Social Forces, vol. 85, no. 2, Dec. 2006, pp. 667–695, https://doi.org/10.1353/sof.2007.0002.

Childress, S., et al. "A Qualitative Study of Adaptation Challenges of Ukrainian Refugees in the United States." Children and Youth Services Review, vol. 169, Nov. 2024, p. 108039, https://doi.org/10.1016/j.childyouth.2024.108039.

Corrales, Javier, and Michael Penfold. Dragon in the Tropics. Rowman Littlefield, 1 Jan. 2011, https://doi.org/10.5771/9780815705024.

Federal Register. "Implementation of the Uniting for Ukraine Parole Process." federalregister.gov, 6 Apr. 2022, https://www.federalregister.gov/documents/2022/04/27/2022-09087/implementation-of-the-uniting-for-ukraine-parole-process. Accessed 14 July 2026.

Fihel, Agnieszka, et al. "Diaspora at War: Mobilization of the Ukrainian Diaspora in the First 2 Years of the Russian Full-Scale Invasion." Journal of Refugee Studies, vol. 38, no. 3, Feb. 2025, pp. 704–721, https://doi.org/10.1093/jrs/feaf006.

Folsom, Emilie, and Jeanne Batalova. "Ukrainian Immigrants in the United States." Migration Policy Institute, 12 Feb. 2026,

https://www.migrationpolicy.org/journal/spotlight/ukrainian-immigrants-united-states. Accessed 17 July 2026.
Freier, Luisa Feline, and Nicolas Parent. “The Regional Response to the Venezuelan Exodus.” Current History, vol. 118, no. 805, Feb. 2019, pp. 56–61, https://doi.org/10.1525/curh.2019.118.805.56.
Frelick, Bill. “What’s Wrong with Temporary Protected Status and How to Fix It: Exploring a Complementary Protection Regime.” Journal on Migration and Human Security, vol. 8, no. 1, Feb. 2020, pp. 42–53, https://doi.org/10.1177/2331502419901266.
Godinho Nascimento, Helena, et al. “The Nature and Extent of Migrants, Immigrants, and Refugees’ Trust in the Social Institutions of Their Host Countries: A Scoping Review.” Frontiers in Political Science, vol. 7, Sept. 2025, https://doi.org/10.3389/fpos.2025.1573017.
Guelespe, Diana et al. “Examining Afghan Evacuees’ Resettlement: Insights and Lessons for Future Humanitarian Populations.” Urban Institute, Mar. 2024, https://www.urban.org/research/publication/examining-afghan-evacuees-resettlement-insights-and-lessons-future.
Haidari, M. Ashraf. “Policy Brief: Afghan Adjustment Act of 2025 and the Protection Way Forward - Displaced International.” Displaced International, 26 Aug. 2025, https://displacedinternational.org/policy-brief-afghan-adjustment-act-of-2025-and-the-protection-way-forward/. Accessed 14 July 2026.
Hainmueller, Jens, et al. “When Lives Are Put on Hold: Lengthy Asylum Processes Decrease Employment among Refugees.” Science Advances, vol. 2, no. 8, Aug. 2016, p. e1600432, https://doi.org/10.1126/sciadv.1600432.
Hanna, Caroline, "The Impact of Human Rights Abuses and Emigration in Venezuela" (2021). Williams Honors College, Honors Research Projects. 1439. https://ideaexchange.uakron.edu/honors_research_projects/1439
Hayduk, Ron, et al. “Immigrant Engagement in Participatory Budgeting in New York City.” New Political Science, vol. 39, no. 1, Jan. 2017, pp. 76–94, https://doi.org/10.1080/07393148.2017.1278855.
Hou, Feng. “The Resettlement of Vietnamese Refugees across Canada over Three Decades.” Journal of Ethnic and Migration Studies, vol. 47, no. 21, Sept. 2020, pp. 4817–4834, https://doi.org/10.1080/1369183x.2020.1724412.
Hyndman, Jennifer, et al. “Sustaining the Private Sponsorship of Resettled Refugees in Canada.” Frontiers in Human Dynamics, vol. 3, May 2021, https://doi.org/10.3389/fhumd.2021.625358.
Ibáñez, Ana María, et al. “Life Out of the Shadows: The Impacts of Regularization Programs on the Lives of Forced Migrants.” Journal of the European Economic Association, vol. 23, no. 3, Sept. 2024, pp. 941–982, https://doi.org/10.1093/jeea/jvae044.

---. “Protecting the Vulnerable: How Migrant Regularization Reduces Crime and Empowers Women.” Journal of Development Economics, vol. 179, Feb. 2026, p. 103667, https://doi.org/10.1016/j.jdeveco.2025.103667.

Interviewee 1. Personal Interview with Anonymous Refugee, Conducted by Long Nguyen, 5 Jul. 2025, Houston, Texas.

Islam, Sazzad. “PUBLIC FINANCE AND POLICY EFFECTIVENESS A REVIEW OF PARTICIPATORY BUDGETING IN LOCAL GOVERNANCE SYSTEMS.” Journal of Sustainable Development and Policy, vol. 01, no. 01, Mar. 2025, pp. 115–143, https://doi.org/10.63125/p3p09p46.

Jacome, Sophia. “Rally Brings Thousands to D.C. Ahead of Third Anniversary of Full-Scale Invasion of Ukraine.” The Georgetown Voice, 28 Feb. 2025, https://georgetownvoice.com/2025/02/28/rally-brings-thousands-to-d-c-ahead-of-third-anniversary-of-full-scale-invasion-of-ukraine-by-sophia-jacome/. Accessed 16 July 2026.

Kozachenko, Ivan. “Diasporic Mobilisation and Repositioning: The Ukrainian World Congress’ Responses to Critical Junctures and Events.” Central and Eastern European Migration Review, Nov. 2025, https://doi.org/10.54667/ceemr.2025.27.

Krawchenko, Tamara. “Ukraine’s Decentralisation Reforms and the Path to Reconstruction, Recovery and European Integration.” LSE Public Policy Review, vol. 3, no. 1, Sept. 2023, https://doi.org/10.31389/lseppr.87.

La Corte, Matthew. “10 Takeaways from New Data on Who Signed up to Sponsor Displaced Ukrainians.” Niskanen Center, 6 Oct. 2022, https://www.niskanencenter.org/10-takeaways-from-new-data-who-signed-up-to-sponsor-displaced-ukrainians/.

Leblang, David et al. “Temporary Protected Status and Immigration to the United States.” Cato Institute, 23 Jan. 2019, https://www.cato.org/publications/research-briefs-economic-policy/temporary-protected-status-immigration-united-states.

Levi, Margaret, and Laura Stoker. “Political Trust and Trustworthiness.” Annual Review of Political Science, vol. 3, no. 1, June 2000, pp. 475–507, https://doi.org/10.1146/annurev.polisci.3.1.475.

Media Center Ukraine. “Since Start of Full-Scale Invasion, over USD 100 Million in Aid Collected by Razom for Ukraine Charity Organization - Media Center Ukraine.” Media Center Ukraine, 23 Aug. 2023, https://mediacenter.org.ua/since-start-of-full-scale-invasion-over-usd-100-million-in-aid-collected-by-razom-for-ukraine-charity-organization/. Accessed 16 July 2026.

Menjívar, Cecilia. “Liminal Legality: Salvadoran and Guatemalan Immigrants’ Lives in the United States.” American Journal of Sociology, vol. 111, no. 4, Jan. 2006, pp. 999–1037, https://doi.org/10.1086/499509.

Migliorini, Laura, et al. “Community Resilience among Ukrainian Refugees: What Is the Role of the Host Community in Recovery from Forced Migration?” Frontiers in Psychiatry, vol. 14, Aug. 2023, p. 1206304, https://doi.org/10.3389/fpsyt.2023.1206304.

Mishler, William, and Richard Rose. “What Are the Origins of Political Trust?” Comparative Political Studies, vol. 34, no. 1, Feb. 2001, pp. 30–62, https://doi.org/10.1177/0010414001034001002.

Montalvo, Julian, and Jeanne Batalova. “Afghan Immigrants in the United States.” Migration Policy Institute, 15 Feb. 2024, https://www.migrationpolicy.org/journal/spotlight/afghan-immigrants-united-states. Accessed 14 July 2026.

Mossaad, Nadwa, et al. “Determinants of Refugee Naturalization in the United States.” Proceedings of the National Academy of Sciences, vol. 115, no. 37, Aug. 2018, pp. 9175–9180, https://doi.org/10.1073/pnas.1802711115.

NBC6. “A Red Wave Swept Across Miami-Dade Elections. We Break Down Results of Key Races.” NBC 6 South Florida, 6 Nov. 2024, https://www.nbcmiami.com/news/local/a-red-wave-swept-across-miami-dade-county-elections-we-break-down-results-of-key-races/3465284/. Accessed 14 July 2026.

Noe-Bustamante, Luis. “Latinos Make up Record 17% of Florida Registered Voters In 2020.” Pew Research Center, 19 Oct. 2020, https://www.pewresearch.org/short-reads/2020/10/19/latinos-make-up-record-17-of-florida-registered-voters-in-2020/. Accessed 14 July 2026.

Office of Inspector General. “DHS Has a Fragmented Process for Identifying and Resolving Derogatory Information for Operation Allies Welcome Parolees | Office of Inspector General.” Dhs.Gov, 6 May 2024, https://www.oig.dhs.gov/reports/2024/dhs-has-fragmented-process-identifying-and-resolving-derogatory-information-operation-allies-welcome-parolees/oig-24-24-may24. Accessed 15 July 2026.

Our Impact — Afghan-American Community Organization. “Afghan-American Community Organization.” Afghan-American Community Organization, 2025, https://aa-co.org/our-impact. Accessed 14 July 2026.

Padgett, Tim. “FIU Poll Shows Florida Venezuelans, Once Pro-Trump, Soured and Polarized About Him Now.” WLRN, 5 May 2025, https://www.wlrn.org/government-politics/2025-05-05/florida-venezuelans-trump-support-fiu-poll. Accessed 14 July 2026.

Parness, Ayelet. “Asylum Backlog Presents Anguish, Uncertainty for Seekers.” HIAS, 4 Apr. 2024, https://hias.org/news/asylum-backlog-presents-anguish-uncertainty-seekers/. Accessed 14 July 2026.

Popov, Bohdan. “Diaspora Diplomacy: How Ukrainian Americans Strengthen U.S.-Ukraine Ties.” Thegaze.Media, 2025,

https://thegaze.media/news/diaspora-diplomacy-how-ukrainian-americans-strengthen-us-ukraine-ties. Accessed 14 July 2026.
Prantl, Janine, "Community Sponsorships For Refugees And Other Forced Migrants: Learning From Outside And Inside The United States" (January 24, 2023). Georgetown Immigration Law Journal (forthcoming), Available at SSRN: https://ssrn.com/abstract=4328750
Röder, Antje, and Peter Mühlau. "Low Expectations or Different Evaluations: What Explains Immigrants' High Levels of Trust in Host-Country Institutions?" Journal of Ethnic and Migration Studies, vol. 38, no. 5, May 2012, pp. 777–792, https://doi.org/10.1080/1369183x.2012.667986.
Rojas, Daniel, et al. "Protecting Irregular Migrants: Evidence from Colombia." Migration Studies, vol. 13, no. 4, Sept. 2025, https://doi.org/10.1093/migration/mnaf046.
Rubin, Barnett R. "Crafting a Constitution for Afghanistan." Journal of Democracy, vol. 15, no. 3, July 2004, pp. 5–19, https://doi.org/10.1353/jod.2004.0051.
Ruiz Sportmann, Anne Stefanie, and Itay Greenspan. "Relational Interactions Between Immigrant and Native-Born Volunteers: Trust-Building and Integration or Suspicion and Conflict?" VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations, vol. 30, no. 5, Mar. 2019, pp. 932–946, https://doi.org/10.1007/s11266-019-00108-5.
Shapovalova, Natalia, and Olga Burlyuk, editors. Civil Society in Post-Euromaidan Ukraine : From Revolution to Consolidation. Vol. 193, ibidem Press, 2018.
Sõukand, Kaspar. "Ukrainian Diaspora in the U.S. And Its Political Effects." Transatlantic Dialogue Center, 22 Jan. 2026, https://tdcenter.org/2026/01/22/ukrainian-diaspora-in-the-u-s-and-its-political-effects/. Accessed 14 July 2026.
Thomson, Viktoriya. "From the Orange Revolution to Euromaidan? Political Culture of Ukrainian Society from 2004 to 2014." 2019. Carleton University, https://doi.org/10.22215/etd/2019-13437.
Turper, Sedef, and Kees Aarts. "Political Trust and Sophistication: Taking Measurement Seriously." Social Indicators Research, vol. 130, no. 1, Nov. 2015, pp. 415–434, https://doi.org/10.1007/s11205-015-1182-4.
Tyler, Tom R. "Restorative Justice and Procedural Justice: Dealing with Rule Breaking." Journal of Social Issues, vol. 62, no. 2, May 2006, pp. 307–326, https://doi.org/10.1111/j.1540-4560.2006.00452.x.
U.S. Citizenship and Immigration Services. "Certain Afghan and Ukrainian Parolees Are Employment Authorized Incident to Parole | USCIS." Uscis, 21 Nov. 2022, https://www.uscis.gov/archive/certain-afghan-and-ukrainian-parolees-are-employment-authorized-incident-to-parole. Accessed 14 July 2026.
---. "DHS Previews Federal Register Notice Extending and Separately Redesignating Venezuela for Temporary Protected Status | USCIS." Uscis, 29 Sept. 2023,

https://www.uscis.gov/newsroom/news-releases/dhs-previews-federal-register-notice-extending-and-separately-redesignating-venezuela-for-temporary. Accessed 14 July 2026.
---. “DHS Publishes Federal Register Notice Extending Temporary Protected Status for Venezuela | USCIS.” Uscis, 15 Jan. 2025, https://www.uscis.gov/archive/dhs-publishes-federal-register-notice-extending-temporary-protected-status-for-venezuela. Accessed 14 July 2026.
---. “Update: Supreme Court Order for TPS Venezuela | USCIS.” Uscis, 22 June 2026, https://www.uscis.gov/i-9-central/form-i-9-related-news/update-supreme-court-order-for-tps-venezuela-0. Accessed 14 July 2026.
U.S. Customs and Border Protection. “CBP Releases August 2024 Monthly Update.” U.S. Customs and Border Protection, 16 Sept. 2024, https://www.cbp.gov/newsroom/national-media-release/cbp-releases-august-2024-monthly-update. Accessed 14 July 2026.
Ukraine Immigration Task Force. “Uniting for Ukraine: A Four-Year Retrospective – Ukraine Immigration Task Force.” Ukrainetaskforce.Org, 20 May 2026, https://ukrainetaskforce.org/uniting-for-ukraine-a-four-year-retrospective/. Accessed 14 July 2026.
UNHCR. “Afghanistan Refugee Crisis Explained.” The UN Refugee Agency, 29 July 2024, https://www.unrefugees.org/news/afghanistan-refugee-crisis-explained/. Accessed 17 July 2026.
USA for UNHCR. “Ukraine Refugee Crisis: Aid, Statistics and News | USA for Unhcr.” Unrefugees.Org, UNHCR, 2025, https://www.unrefugees.org/emergencies/ukraine/. Accessed 17 July 2026.
Verba, Sidney, et al. Voice and Equality: Civic Voluntarism in American Politics. Harvard University Press, 1995. JSTOR, https://doi.org/10.2307/j.ctv1pnc1k7. Accessed 13 July 2026.
Waldron, William. “How Can We Save Venezuela’s Economy?” SSRN Electronic Journal, Center for Economic Studies Ifo Institute, 1 Jan. 2026, https://doi.org/10.2139/ssrn.5994415.
Walker, Nigel. “Conflict in Ukraine: A Timeline (2014 – Eve of 2022 Invasion).” House of Commons Library, 22 Aug. 2023, https://commonslibrary.parliament.uk/research-briefings/cbp-9476/. Accessed 18 July 2026.
Weinbaum, Marvin, and Naade Ali. “Thousands of Once Protected Afghan Refugees in the United States Face Deportation - Middle East Institute.” Middle East Institute, 20 Nov. 2025, https://mei.edu/publication/thousands-once-protected-afghan-refugees-united-states-face-deportation/. Accessed 15 July 2026

Safe Space By Aanya Yadawad

Abstract

The project Safe Space is an AI-based chatbot designed to help users identify whether they may be experiencing symptoms of depression or anxiety. Safe Space asks users a total of five questions. These questions are intended to help the model identify whether the user is experiencing symptoms of anxiety or depression. I used data from a Reddit dataset. I extracted relevant keywords from the Reddit dataset in order to classify symptoms as either depression or anxiety. Afterwards, I developed a chatbot website that sends user responses to a backend model and returns a recommendation rather than a formal diagnosis.

Introduction

The most common mental illnesses in the United States are anxiety disorders ("Anxiety Disorders"). Individuals who have anxiety disorders often experience extreme and continuous fear and worry. Common and recognizable symptoms of anxiety include feeling nervous and tense, having an increased heart rate, hyperventilation (breathing rapidly), sweating, and trembling ("Anxiety disorders"). Anxiety disorders affect 19.1% of adults in the United States, which is around 40 million people. Although anxiety disorders are highly treatable, only 1 in 4 people receive treatment for their illness. This means that only 27.6% of people with anxiety disorders receive treatment ("Anxiety Disorders").

On the other hand, depression is a condition characterized by extreme sadness that can interfere with everyday life. Common symptoms of depression include feeling anxious or sad for long periods of time, losing interest in activities that were once enjoyable, or experiencing thoughts of suicide or self-harm. Approximately 1 out of every 6 adults will experience depression at some point in their lives. Around 16 million American adults are affected by depression every year. Many people who experience depression also have a high likelihood of experiencing anxiety disorders. Both depression and anxiety disorders often overlap ("Mental Health"). With all this data, the question remains: why do so many people continue to suffer today?

There is not one specific reason for this, but one contributing factor is that many people are not seeking treatment for their conditions. According to the American Psychological Association, many Americans have reported that their mental health needs have not been met. Affordability remains one of the top reasons people do not seek help, but other barriers are also preventing individuals from receiving treatment. In 2018, 39% of people reported that they could not afford treatment for their mental illness. Other major reasons they did not seek help included believing they could manage their illness without treatment, not knowing where to go for services, and not having enough time to seek treatment (Conroy et al.). Some obvious limitations stem from professional evaluations not being immediately accessible, screening tools lacking interactivity, users hesitating to seek professional help due to stigma or inconvenience, and existing tools failing to provide

personalized recommendations. Therefore, having accessible and interactive tools that can provide preliminary guidance regarding symptoms while encouraging users to seek professional care when appropriate is crucial.

In order to address this issue, I created Safe Space. Safe Space was designed to combat these issues and provide users with guidance for their mental health. What I built from Safe Space was an AI-based chatbot for depression and anxiety recommendations. I trained a machine learning model using a Reddit dataset. I took out keywords and language patterns associated with anxiety and depression. Then, I created a web interface connected to a backend classification model. I was able to create a web page that collected user responses and responded with a recommendation of support.

Related Work

Artificial intelligence (AI) has been an important tool for developing mental health chatbots that provide immediate and accessible emotional support. These systems rely on natural language processing to understand user responses and respond in a supportive and conversational way. Over the years, there have been several mental health chatbots made to help people. One of the most widely used mental health chatbots is Wysa. Communicating with Wysa is free and it includes tools that are offered with conversations (Wysa). Another prominent mental health chatbot includes Youper. Youper uses artificial intelligence to help you track your mood, emotional well-being, and it even helps write down your thoughts in an organized journal. It also emphasizes some of its limitations as an AI chatbot. It might not be able to recognize all feelings of pain or emotional distress. Youper also does not serve as a substitute for professional mental health care (Youper). Ash is also an advanced chatbot. Ash uses AI tools that were built using data from the online web. Ash is an expert in human psychology and behavior. It is trained on huge-scale of proprietary expert data (Talk to Ash). All of these chatbots have one thing in common: they are free to use (except Youper which shifted to a subscription model) and they are extremely easy to access.

These chatbots have combated one of the most common reasons people do not seek help for their mental health: affordability. All of these mental health chatbots are also very advanced, unlike Safe Space. Safe Space has quite a few differences. Firstly, Safe Space only focuses on depression and anxiety. The other chatbots serve a broader range of mental health tools and emotional tracking features that Safe Space does not. Safe Space is more limited in scope. Also, Safe Space, unlike Ash, only uses a specific Reddit dataset for its information. The information Safe Space uses is very limited compared to more advanced mental health chatbots such as Wysa, Youper, and Ash. However, if there is one thing Safe Space, Wysa, Ash, and Youper have in common, it's that they all do not provide official diagnoses and cannot provide professional mental health care, an important note for those seeking real treatment for their illness.

Methods

The Reddit dataset was first saved into a file using Google Colab for further analysis. To access the dataset, the local computer was connected to the Colab environment by importing the Google Drive module and mounting the drive using the commands from google.colab import drive and drive.mount('/content/drive'). The dataset consisted of 2,390 rows, with each entry labeled as either depression (1) or anxiety (4). A pretrained embedding model was then imported from the SentenceTransformers library to convert Reddit posts into vector representations using the "all-MiniLM-L6-v2" model. This conversion was necessary because machine learning models cannot process raw text directly and instead require numerical input. To test the transformation process, embeddings were generated for the first ten Reddit posts, and a function was created to store these embeddings in a list for easier handling. The list was subsequently converted into a NumPy array because the machine learning model required array input rather than a list.

A logistic regression model was implemented using scikit-learn to classify Reddit posts as either depression (1) or anxiety (4). An initial prediction was performed on a single Reddit post, specifically the first post in the dataset, which the model classified as anxiety (4). Model accuracy was calculated using the formula Accuracy = np.mean(prediction == y) to evaluate performance. The initial accuracy was approximately 0.5, indicating that the model required further improvements.

Several optimizations were then introduced to improve performance and efficiency. The embedding model initialization was moved to the beginning of the code to prevent repeated loading and reduce computational overhead. Rows containing missing or invalid (NaN) values were removed to improve data quality. The dataset was split into training and testing subsets using a train-test split method, with the variables divided using train_test_split (X, y, test_size=0.33, random_state=42). This allowed the model to be trained on one portion of the data and evaluated on another, providing a more reliable measure of accuracy. In addition, the sample size was increased to 1,000 data points to improve the robustness of the model, and a more complex neural network model was implemented to achieve better predictive performance. Following model development, a web-based chatbot interface was created using Visual Studio Code. HTML paragraph tags (<p>) were used to display introductory text on the webpage, including a greeting message for users. A form containing a textarea input field, placeholder prompt, and submit button was added to allow users to enter responses. A list of diagnostic questions was defined within the front-end code to guide user interaction, and conditional logic was implemented so that the final question displayed a **Submit** button instead of a **Next** button. A Python file named model.py was created to handle diagnostic predictions, while a virtual environment was established using python3 -m venv chatbox_env and activated using source chatbox_env/bin/activate to manage project dependencies and install required libraries.
To connect the front end and back end, an Axios function was implemented to send user responses to the server for processing. A data pipeline was established to transfer user input from the web interface to the machine learning model, where a variable named diagnosis stored the model's recommendation. After processing the responses, the model generated a diagnosis

message recommending that the user seek a mental health professional for depression. Finally, once the diagnosis was produced, the user interface was updated to remove the previous elements and display only the final recommendation.

Algorithm 1: Mental Health Classification

Input: response_text
Output: diagnosis_message

BEGIN
embedding_model ← Load Sentence Transformer
embedding ← Encode(response_text)

classifier ← Load Trained Model
prediction ← classifier.Predict(embedding)

IF prediction = 1 THEN
diagnosis ← "depression"
ELSE
diagnosis ← "anxiety"
END IF

RETURN "I think you should seek a mental health professional for " + diagnosis
END

Results

I evaluated the Safe Space model using training and testing its accuracy. I put in different answers to the questions and identified whether it accurately guessed the symptoms as either depression or anxiety. The model achieved an overall accuracy of 50%, demonstrating its fluctuating ability to correctly identify depression and anxiety symptoms. The results signaled that the chatbot was not fully capable in providing the most accurate assistance for preliminary mental health screening.

Model Accuracy Graph

Inputs:	Accuracy of Outputs:
User input 1	Correct
User input 2	Incorrect

User input 3	Incorrect
User input 4	Incorrect
User input 5	Correct
User input 6	Correct

Discussion

I believe my diagnostic model could have several improvements. I believe this is a very good project for learning the basics of coding and having a model that is working. I believe that this model is not very accurate and mainly looks for key words in user input as its main source of information. It can be accurate for some people but inaccurate for others. I think labels are very noisy. People who are active in Reddit communities talking about depression might not actually be depressed. A person might post there because they are simply having a rough week or feel lonely. It does not guarantee they have a medically diagnosed depressive disorder. Furthermore, there is no metadata. Richer datasets normally have additional contextual information such as posting date, subreddit name, number of upvotes, etc. There is also an obvious bias in the dataset. Reddit users skew younger, more Western, and more male. Therefore, there is not enough diversity in the dataset's user population. In the future, I would improve this project with a better dataset. I would also make the website more colorful and engaging. For someone seeking mental health help, the website seems very plain and boring. The color choices were not that comforting.

Conclusion

In conclusion, artificial intelligence and machine learning were used to create *Safe Space* in order to provide people with accessible mental health guidance in depression and anxiety. The Reddit-based dataset and natural language processing techniques helped me build a chatbot capable of analyzing user input and providing supportive recommendations. Despite the model having limitations in accuracy, dataset diversity, and clinical reliability, this project was successful for fundamental coding, machine learning, text classification, and web development. Most importantly, *Safe Space* was created with the purpose of encouraging users to seek professional help and make mental health support feel more accessible, private, and approachable.

Works Cited

"Anxiety disorders." Mayo Clinic, 29 July 2025, www.mayoclinic.org/diseases-conditions/anxiety/symptoms-causes/syc-20350961. Accessed 10 May 2026.

"Anxiety Disorders - Facts & Statistics." Anxiety and Depression Association of America, 30 Mar. 2026, adaa.org/understanding-anxiety/facts-statistics. Accessed 10 May 2026.

Conroy, Jessica, et al. "Why people aren't getting the care they need." American Psychological Association, 1 July 2020, www.apa.org/monitor/2020/07/datapoint-care. Accessed 10 May 2026.

Ghoshal, Neel. "Reddit Mental Health Data." 2023. Kaggle, www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data. Accessed 10 May 2026.

"Mental Health Conditions: Depression and Anxiety." CDC, 13 Oct. 2023, www.cdc.gov/tobacco/campaign/tips/diseases/depression-anxiety.html. Accessed 10 May 2026.

"Mental Health Disorder Statistics." John Hopkins Medicine, www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics. Accessed 10 May 2026.

Talk to Ash. Talk to Ash, www.talktoash.com/. Accessed 25 June 2026.

Wysa. Wysa Ltd., www.wysa.com/faq?utm_source=chatgpt.com.

Youper. Web Page, www.youper.ai/. Accessed 25 June 2026.

Too Much, Too Soon: Cultural Lag and Adolescent Exposure to Digital Pornography
By Avery Solomon

Abstract

Over the last thirty years, the rapid increase in digital access to pornography has had a profound effect on the sexual, moral, and physical well-being of adolescents. Social institutions, including the legal, educational, and familial systems, have made various attempts to adapt to this ever-changing landscape. In this paper I examine these attempts and seek to assess the extent to which they have been successful. Through this research I concluded that the rapid expansion of internet pornography demonstrates the phenomenon of cultural lag, as technological advancements have far outpaced the ability of legal systems, educational institutions, families, and digital platforms to adapt effectively. This disconnect has produced conditions of anomie in which children and adolescents encounter adult sexual norms without adequate social guidance, resulting in developmental and social consequences that existing institutions remain ill-equipped to address.

Introduction

Childhood has always been shaped by the boundaries a society draws between what is revealed and what is withheld. Pornography, understood as the depiction of erotic behavior intended to arouse, is not a novel feature of modern society. From the frescoes of Pompeii and the sculptures of ancient Greece to the erotic poetry of early modern Europe, erotic imagery can be traced across culture and history. However, these early depictions of sensuality were embedded within specific cultural, religious, and spatial contexts that shaped how such material was interpreted and regulated (Williams). The concept of pornography as a distinct social category emerged much later, alongside the development of mass print technologies in the eighteenth and nineteenth centuries, when erotic pamphlets, novels, and illustrations became more widely distributed and increasingly subject to moral and legal scrutiny (Garton). Throughout the twentieth century, technological innovation continued to reshape the accessibility of sexually explicit material. The introduction of film allowed pornography to circulate in underground theaters and private venues, leading to the sexual revolution of the 1960s and 1970s that briefly brought explicit films into public cultural discourse during what has been described as the "Golden Age of Porn" (Williams). Later, the rise of home video technologies and subscription cable television in the 1980s marked a significant shift, relocating pornography from public or semi-public spaces into the privacy of the home. Scholars argue that this transition altered the mechanisms through which pornography was informally regulated, allowing for individual control over consumption (Attwood). However, the widespread adoption of the internet represents a more profound rupture in the distribution of pornography. Beginning in the late 1990s and accelerating rapidly with the spread of broadband internet and smartphones, pornography became instantaneous, portable, and largely detached from spatial and social constraints. Online platforms not only expanded access but also introduced algorithmic systems

that personalize and intensify content exposure, fundamentally altering how sexually explicit material is encountered and consumed (Paasonen; Döring). Unlike earlier forms of pornography, which required deliberate pursuit and were only accessible via physical media, internet pornography exists in a state of constant proximity, embedded within the everyday digital environments of users. This transformation has had particularly significant implications for children and adolescents. Empirical research consistently demonstrates that a majority of young people encounter online pornography during adolescence, often unintentionally and at increasingly early ages (Sabina et al.). Yet the social institutions historically responsible for mediating sexual knowledge —law, education, family, and community — have struggled to adapt to this new technological landscape. Legal institutions governing pornography remain rooted in assumptions about physical distribution and adult-only spaces, while educational institutions frequently avoid direct engagement with explicit media due to political controversy and outdated models of sex education (Attwood).

Cultural lag, as first described by William Ogburn, is a sociological concept that identifies periods where societal norms fail to catch up to a newly introduced technology. In the context of internet pornography, my hypothesis is that technological innovation has outpaced the ability of social institutions to establish coherent norms surrounding exposure, interpretation, and protection.

Literature Review

Existing scholarship traces pornography's shift from a spatially and socially regulated medium to a digitally ubiquitous one. Williams situates the mainstreaming of explicit film within the sexual revolution of the 1970s, examining the effect that the digital revolution had on social-sexual perceptions of women. Her findings present a conclusion that hard-core digital pornography only perpetuates a “phallic” perspective on pleasure while presenting women as the method to obtain it. Attwood argues that home video and cable technologies moved consumption into private, individually regulated spaces during the 1980s, leading to social “pornographication,” making it easier to push narratives about women to a diversified audience. She argues that after the progress made by the feminist movement in the 1970s, digital pornography became a method to disguise a new form of sexism. Paasonen and Doring extend this trajectory into the internet era, describing how algorithmic personalization has intensified and individualized exposure in ways earlier media could not. Their research moves into an analysis of how this new technology has begun to affect both adults and adolescents, marking some of the first mentions of the subject in scientific writing.

Empirical work on adolescents specifically, most notably Sabina and colleagues' foundational study on the nature and dynamics of youth exposure, and Andrie and colleagues' cross-national survey of European adolescents, establishes that exposure is now widespread, frequently unintentional, and occurring at increasingly early ages. Their data quantifies the detrimental effects of online pornography on youth as it grows in accessibility and awareness. Legal scholarship documenting cases like Reno v. ACLU and the UK's Digital Economy Act

shows the repeated failure of obscenity based regulatory frameworks to adapt to a decentralized, borderless medium. As courts around the world struggle to regulate the technology, it becomes clearer how lacking our regulatory systems are.

Discussion

Before the digital era, legal regulation operated as an effective — if imperfect—mechanism for controlling access to pornographic material. This was because the system was embedded within a broader network of social norms and shared assumptions about visibility, space, and age-appropriate boundaries. In the United States, federal obscenity law is grounded in the Miller test, a three-pronged standard articulated by the Supreme Court in *Miller v. California* (1973) that evaluates material according to whether it appeals to prurient interests, depicts sexual conduct in a blatantly offensive way, and lacks serious literary, artistic, political, or scientific value (Columbia Global Freedom of Expression). This framework assumes that sexually explicit material circulates through identifiable channels—bookstores, theaters, or mailed publications—where access can be regulated by physical observation. These legal restrictions were reinforced by informal social mechanisms: shared media spaces, reputational consequences for public consumption, and adult-only settings that limited exposure, especially among youth.

Attempts to extend traditional legal controls into the digital environment exposed the limits of these assumptions. Early U.S. legislative efforts to regulate sexually explicit material online, such as provisions of the Communications Decency Act (CDA) of 1996 and the Child Online Protection Act (COPA), sought to implement existing obscenity standards into online spaces. However, the CDA's anti-indecency provisions were struck down by the Supreme Court in *Reno v. American Civil Liberties Union* (1997) as overly broad and incompatible with First Amendment protections, along with COPA which was similarly repealed on constitutional grounds ("Round Two: Cyber-Censorship Returns to Supreme Court | American Civil Liberties Union"). Legal commentators explain that judges found these statutes unenforceable in part because they relied on categories of obscenity that function with the assumption of clearly bound media and definable audiences, neither of which accurately describes the decentralized, borderless nature of the internet (Klosek). The consumption of pornography became private and individualized, occurring on personal devices rather than in observable spaces. Legal rules that once depended on the visibility of media networks lost much of their regulatory force in an environment characterized by anonymity.

The United Kingdom's legislative attempt to mandate age verification under Part 3 of the Digital Economy Act 2017 further illustrates this point: although the statute directed the British Board of Film Classification to enforce age checks, implementation was delayed and ultimately rescinded due to concerns about privacy, enforcement mechanisms, and technological feasibility (Woodhouse). Scholars analyzing these reforms emphasize that age verification requirements failed to account for technical limitations and the digital literacy of adolescents ("Online Privacy Regulation"). Legal systems continue to rely on the physical existence of pornography which no

longer applies in the digital age. While technology governing the production and distribution of pornography has evolved rapidly, the legal and social systems designed to regulate it remain stunted in earlier media environments. The result is a widening gap between authority and digital reality, one that leaves exposure increasingly unregulated despite the persistence of formal legal standards.

The expansion of digital technology has transformed pornography from a regulated, spatially bounded medium into a ubiquitous feature of adolescent digital life, fundamentally altering patterns of exposure and socialization. Whereas prior encounters with sexually explicit material were limited by physical access, contemporary adolescents navigate an online environment in which pornography is algorithmically promoted, instantly accessible, and largely detached from institutional oversight. As a result, exposure has become widespread and increasingly normalized: cross-national research indicates that 59 % of adolescents report exposure to online pornography, with 24 % encountering it at least weekly, with heavier internet use increasing the frequency of exposure (Andrie et al.). Importantly, exposure now occurs earlier than in previous generations. European surveys report that 95 % of boys and 70% of girls encounter pornography during adolescence, with an average initial exposure at approximately 11.7 years (Sabina et al.). Recent meta-analysis indicates that exposure to sexually explicit content significantly increases the likelihood of problematic sexual behaviors, including multiple sexual partners and unprotected sex, with odds ratios ranging from 1.82 for non-violent content to 2.52 for violent content compared to unexposed peers (Paulus et al. 654; fasshera). In addition to unsafe sex practices, longitudinal research shows that adolescents who view pornography frequently are more likely to exhibit increases in sexually harassing behaviors over time, particularly among male youth, suggesting that consumption may reinforce gendered patterns of aggression and normative understanding of sexual conduct (Waterman et al. 1612). Yet the consequences are not solely behavioral. Researchers also identify internalized effects such as confusion surrounding consent and intimacy, premature adoption of adult sexual norms, and feelings of shame and secrecy, which can distort adolescents' developmental understandings of relationships and selfhood (Andrie et al.). Technological systems have rapidly reshaped sexual socialization, while institutions have failed to evolve at a comparable pace, leaving adolescents to interpret adult sexual behavior in the absence of adequate social guidance.

In response to the rapid expansion of online pornography and rising concern over adolescent exposure, a range of legal, technological, and familial interventions have been introduced to mitigate potential harms. However, research suggests that these efforts, while well-intentioned, have remained largely insufficient due to their irregularity, limited capacity for enforcement, and misalignment with adolescents' actual digital behavior. As a result, these institutions have struggled to regulate pornography in ways that meaningfully address the sociological consequences of early and pervasive exposure. Legal approaches have primarily focused on age-based restrictions and obscenity law, but these sanctions were once again developed for the restriction of physical media rather than universal online platforms.

In the United States, constitutional protections for free speech have limited the federal government's ability to regulate online sexual content, effectively shifting responsibility toward parents and private platforms. Internationally, some governments have turned to mandatory age-verification systems, particularly in parts of Europe and the United Kingdom. However, research and policy analysis indicate that such systems are often easily circumvented, inconsistently enforced, and accompanied by significant privacy concerns, as they may require users to submit personal identification to third-party services (Woodhull Freedom Foundation). These limitations highlight the difficulty of applying traditional laws in the face of significant cultural change. Technological and platform-based solutions have likewise proven inadequate. Social media and content-hosting platforms have implemented moderation policies, warning labels, and algorithmic filters intended to limit minors' exposure to explicit material, yet studies examining content moderation on major video platforms have found that accounts registered to younger users encounter harmful or sexually explicit material at high rates despite these moderation systems (Eltaher et al.).

At the familial level, parental control software and internet filtering tools are frequently promoted as primary solutions to adolescent pornography exposure. However, research from the Oxford Internet Institute reveals that to prevent one adolescent exposure, between 17 to 77 households would need to implement digital filters (Przybylski). Reliance on parental controls places disproportionate responsibility on families while failing to account for adolescents' technological fluency and mobile device access. Surveys have shown that both parents and adolescents view age-verification tools and filters as largely ineffective against intentional access to explicit material, emphasizing that technology alone cannot substitute for meaningful guidance and education (Turvey et al.).

Recommendations

Addressing the cultural lag evident in digital pornography requires a shift away from prohibition-centered strategies toward institutional reform. Rather than attempting to eliminate exposure altogether—a goal increasingly recognized as unrealistic—schools, public health agencies, and digital platforms must focus on equipping adolescents with the knowledge necessary to contextualize sexually explicit content. With the integration of comprehensive digital sexual literacy into educational curricula, emphasizing consent, relational dynamics, power, and the distinction between scripted and real intimacy, adolescents will have the knowledge needed to more safely encounter explicit material. High school health classes would be expanded to cover the psychological and physiological effects of digital pornography, and the ways in which it can affect relationships. Discussions of consent would also include the ways in which pornography portrays sexual acts in an unrealistic and demeaning manner. By normalizing discussion of pornography and the surrounding subjects, the novelty is removed, which helps to prevent the initial encounter with pornography that may lead to addiction. Surrounding the issue of adolescent pornography exposure is a general societal unease regarding sexual education. It is imperative that budding teenagers understand the realities of intimacy, and how normal it is.

Conclusion

Adolescents, interacting with pornographic materials earlier and more frequently than ever before, are compelled to navigate intimacy, consent, and relational norms largely on their own. Technology evolves at a pace that leaves educational, legal, and parental institutions perpetually a step behind. In a world where technology accelerates faster than institutions can respond, the resulting lag reveals the societal cost of unchecked technological innovation, where those most affected by adults' innovations are children. As childhood access to technology grows, the consequences of the unregulated internet are imposed on those most impressionable, forever changing how the next generation will view and experience the world. One is compelled to question how an artistic expression of sensuality has come so far from the beauty it intended to convey.

Works Cited

Andrie, Elisabeth K., et al. “Adolescents’ Online Pornography Exposure and Its Relationship to Sociodemographic and Psychopathological Correlates: A Cross-Sectional Study in Six European Countries.” Children, vol. 8, no. 10, Oct. 2021, p. 925. MDPI, https://doi.org/10.3390/children8100925.

Attwood, Feona. “Sexed up: Theorizing the Sexualization of Culture.” Sexualities, vol. 9, no. 1, Feb. 2006, pp. 77–94. https://doi.org/10.1177/1363460706053336.

Columbia Global Freedom of Expression. “Miller V. California - Global Freedom of Expression.” Global Freedom of Expression, 11 Nov. 2025, globalfreedomofexpression.columbia.edu/cases/miller-v-california.

Döring, Nicola M. “The Internet’s Impact on Sexuality: A Critical Review of 15 Years of Research.” Computers in Human Behavior, vol. 25, no. 5, May 2009, pp. 1089–101. https://doi.org/10.1016/j.chb.2009.04.003.

Eltaher, Fatmaelzahraa, et al. “Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms.” arXiv.org, 16 May 2025, arxiv.org/abs/2505.11160.

Garton, Stephen. Histories of Sexuality. Routledge, 2008.

Klosek, Katherine. Issue Brief: Section 230 of the Communications Decency Act: Research Library Perspectives. 6 July 2021, https://doi.org/10.29242/brief.section230cda2021.

Mubasshera, Hamida. “Pornography Usage During Adolescence: Does It Lead to Risky Sexual Behavior?” Health Economics, vol. 33, no. 8, Mar. 2024, pp. 1682–704. https://doi.org/10.1002/hec.4825.

“Online Privacy Regulation.” Mercatus Center, 6 Dec. 2010, www.mercatus.org/research/policy-briefs/online-privacy-regulation?utm_source=chatgpt.com.

Paasonen, Susanna. “Carnal Resonance.” The MIT Press eBooks, 2011, https://doi.org/10.7551/mitpress/9780262016315.001.0001.

Paulus, Frank W., et al. “The Impact of Internet Pornography on Children and Adolescents: A Systematic Review.” L Encéphale, vol. 50, no. 6, pp. 649–62. https://doi.org/10.1016/j.encep.2023.12.004.

Przybylski, Andrew K., and Victoria Nash. “Internet Filtering and Adolescent Exposure to Online Sexual Material.” Cyberpsychology Behavior and Social Networking, vol. 21, no. 7, July 2018, pp. 405–10. https://doi.org/10.1089/cyber.2017.0466.

“Round Two: Cyber-Censorship Returns to Supreme Court | American Civil Liberties Union.” American Civil Liberties Union, 30 Nov. 2005, www.aclu.org/press-releases/round-two-cyber-censorship-returns-supreme-court.

Sabina, Chiara, et al. “The Nature and Dynamics of Internet Pornography Exposure for Youth.” University of New Hampshire Scholars Repository, scholars.unh.edu/soc_facpub/284.

Turvey, Jake, et al. “Exploring the Feasibility and Acceptability of Technological Interventions to Prevent Adolescents’ Exposure to Online Pornography: Qualitative Research.” JMIR Pediatrics and Parenting, vol. 7, Nov. 2024, p. e58684. https://doi.org/10.2196/58684.

Waterman, Emily A., et al. “Prospective Associations Between Pornography Viewing and Sexual Aggression Among Adolescents.” Journal of Research on Adolescence, vol. 32, no. 4, Mar. 2022, pp. 1612–25. https://doi.org/10.1111/jora.12745.

Williams, Linda. Hard Core: Power, Pleasure, and the “Frenzy of the Visible.” University of California Press, 1999.

Woodhouse, John. “Online Pornography: Age Verification.” UK Parliament, 8551, House of Commons Library, 18 Oct. 2019.

Woodhull Freedom Foundation. “Age Verification Laws Limit Free Speech Online — Woodhull Freedom Foundation.” Woodhull Freedom Foundation, 9 Oct. 2025, www.woodhullfoundation.org/fact-checked/age-verification-limit-speech.

Marksman: An AI Powered Web Application for Intelligence Monitoring People and Objects By Anish Sharma

Abstract

Currently, many common AI platforms, such as ChatGPT or Gemini, function primarily as general-purpose chatbots rather than customizable live-video object classifiers. While object-classification systems exist for applications such as autonomous driving or home-security systems, these models are highly specialized and not designed for general, multi-purpose object recognition. Additionally, many of these solutions are not easily accessible to the general public or underserved communities.

The Marksman created by Anish Sharma addresses this gap by providing a hands-on, customizable video-recognition application. Users can train their own models using personal image datasets to classify custom objects or classes. Built on the Gradio framework, the application allows users to upload images of objects of interest, train an image-recognition model, and perform real-time object prediction through live video. As a result, The Marksman offers an accessible and customizable object-identification solution, distinguishing it from off-the-shelf chatbot-based AI platforms.

Introduction

In today's world, many people have access to powerful AI platforms such as Gemini, Grok, and ChatGPT. However, most of these tools are primarily limited to text-based tasks and do not support real-world applications like long-term video recognition. Although advanced video recognition technologies do exist, they are often not accessible to the general public through commonly used platforms. Additionally, some services that offer video or video analysis charge users per minute of footage processed such as Amazon Rekognition (*Image Recognition Software, ML Image Analysis, and Video Analysis – Amazon Rekognition Pricing – AWS*, 2026), while physical devices with built-in recognition features may not support specific detection classes such as a mailman, neighbor, or friend. As a result, there is a clear need for a video recognition application that can detect user-defined objects and individuals over time.

This problem is important because many physical AI devices or applications with built-in video recognition features are expensive, making them inaccessible to users with limited budgets. Even when such devices are affordable, they often rely on broad detection categories like "motion detected" or "person detected," which do not provide meaningful or actionable information. For instance, a user may want to distinguish between a delivery driver, a family member, or a pet, but most current systems are unable to make these custom distinctions as they do not have the data or classes to detect these events. These limitations reduce the practicality and effectiveness of existing solutions in real-world scenarios. Developing more precise and customizable video recognition tools would allow users to gain greater control, accuracy, and usefulness from AI technology.

The solution to this problem is the development of free or low-cost applications, such as The Marksman, that analyze images or videos to detect specific classes defined by the user. Unlike traditional physical devices, these applications allow users to train the system to recognize exactly what they want, whether it is a pet, a friend, or a mailman. By offering customizable detection, the software removes the constraints of generic categories and provides more accurate and relevant results. Furthermore, these applications are designed to be accessible and affordable, ensuring that users who cannot invest in expensive hardware can still benefit from AI-powered video recognition. In particular, these tasks often require dedicated GPUs such as an NVIDIA GeForce RTX 3060 or 4070. This application compared to other large industrial systems is smaller instead of requiring massive datasets. Overall, solutions like this empower everyday users with long-term monitoring and detection capabilities for personal security, convenience, and other practical real-world applications.

Methods

First, I designed a convolutional neural network (CNN) to classify 64×64 RGB images. To improve generalization and help the model learn more robust features, built-in data augmentation layers are applied at the start of the model. These include random horizontal flips, slight rotations (up to 28.8 degrees or 8%), zooming (up to 10% zooming in and out), translations (up to 8%), and contrast adjustments (*Practical Deep Learning for Coders | Practical Deep Learning for Coders*, n.d.). This allows the model to see multiple variations of each image during training, which reduces overfitting and improves its ability to recognize patterns in new images.

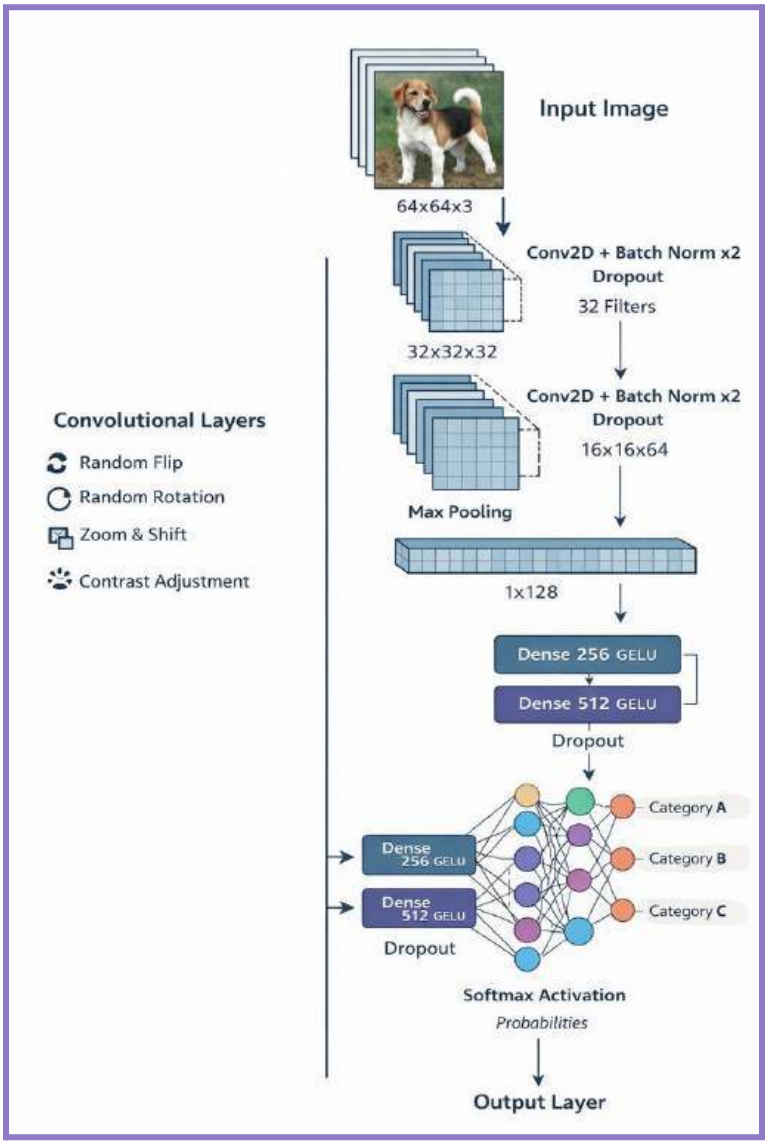
The model then uses three convolutional blocks to extract visual features. Each block applies two convolutional layers with 3×3 filters and same padding, which preserves spatial information while detecting patterns such as edges, shapes, and textures. The number of filters increases in each block (32, then 64, then 128), allowing the network to learn progressively more complex features (TensorFlow, 2019). After each convolution, batch normalization is applied to stabilize the learning process and improve training efficiency. Max pooling layers reduce the spatial dimensions of the image while keeping the most important features, and dropout is used to randomly deactivate some neurons to prevent overfitting and improve generalization.

After feature extraction, the model uses a Global Average Pooling layer to convert the feature maps into a simplified feature vector. This reduces the number of parameters and helps the model focus on the most important learned features (TensorFlow, 2019). The network then passes this information through fully connected Dense layers, which learn higher-level relationships between features and the target categories. Additional Dropout is used to further reduce overfitting.

Finally, the output layer uses a softmax activation function to produce probability scores for each possible class. The class with the highest probability is selected as the model's prediction. Overall, this architecture balances feature extraction, efficiency, and regularization,

allowing the model to learn meaningful patterns while maintaining strong performance on unseen data.

Diagram of Architecture



(Figure describes the Convolutional neural network architecture with data augmentation, convolutional layers, and dense layers for image classification.)

Secondly, I built the Gradio application using Visual Studio Code to prototype it locally. After that, I adapted the code for Hugging Face Spaces, making changes like replacing local file paths with files stored within the application so it would fit the platform's layout, which allowed me to deploy it on public websites and integrate it with the Gradio interface. The Gradio application is built by integrating several deep-learning algorithms and user-interface libraries (Gradio) to create an interactive, web-based computer vision system (Team, 2025). TensorFlow

and Keras are used to design and train a CNN that can classify images based on user-defined labels. To add on, libraries such as NumPy to create arrays/format data, Python Imaging Library to read the images provided by the user, and scikit-learn handle image preprocessing (confusion matrixes). Gradio Blocks are used to structure the interface into tabs, rows, and columns, making the application easy to navigate for tasks such as uploading training data, labeling images, and viewing predictions. The application also allows users to upload multiple images, automatically rename them with custom labels, and download them as a ZIP file for organized dataset creation (Team, 2025). Additionally, the model's performance is visualized directly in the app through a confusion matrix generated with Matplotlib and scikit-learn. Overall, this Gradio application turns a complex deep-learning workflow into an accessible, public-facing web tool that supports training, evaluation, prediction, and live camera input.

Thirdly, the Live Video feature in the application allows users to stream video directly from their webcam and receive real-time predictions from the trained CNN model. Each frame from the webcam is automatically resized to match the input requirements of the model, ensuring that the images or video can be analyzed correctly. The convolutional neural network (CNN) processes each frame of the video stream by extracting hierarchical features learned during training and performing inference to classify the presence of user-defined objects or events. For each input frame, the model outputs a probability distribution across all target classes using a softmax activation function. When the predicted probability for a given class exceeds a predefined confidence threshold, the system registers a detection and triggers a corresponding notification, indicating the identified object or event (2026). This inference pipeline operates continuously on successive frames, enabling real-time analysis of the video feed. By integrating live video input with the trained model, the system supports persistent, automated monitoring without requiring manual image capture or user intervention.

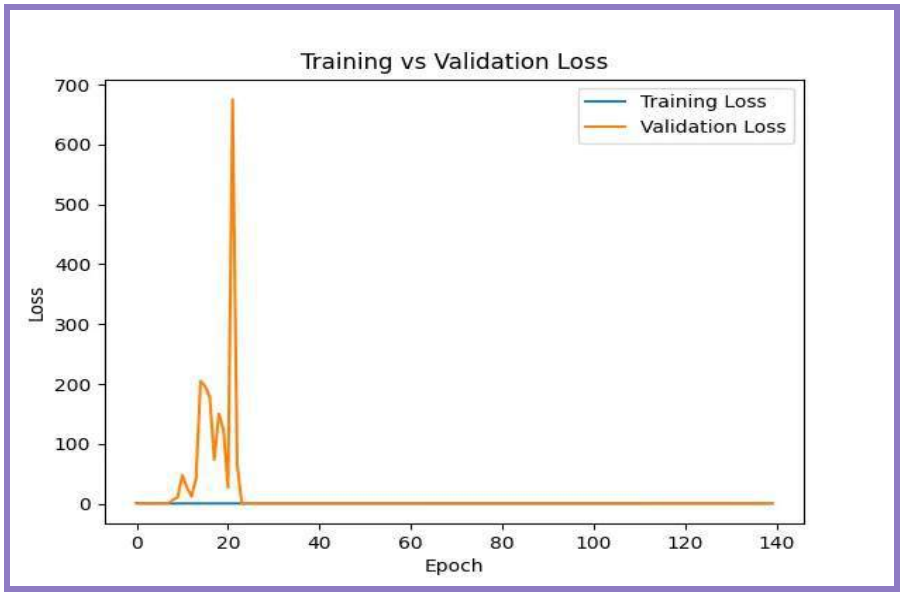
Results

Dataset 1: Humans vs Horses (No. of Files: 1000)

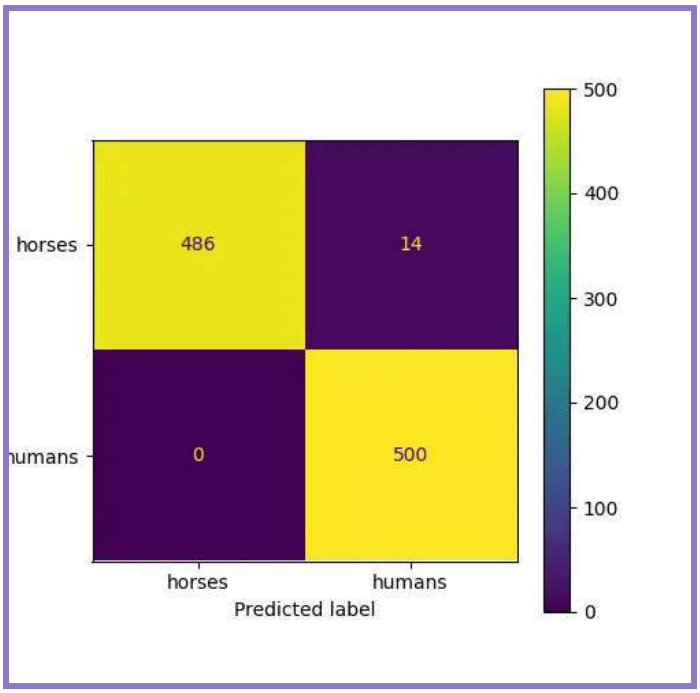
Training Statistics :

- Training Configuration: 5 runs × 100 epochs
- Training Time: 1056 seconds

Performance : Accuracy: 96%, Loss: 0.09



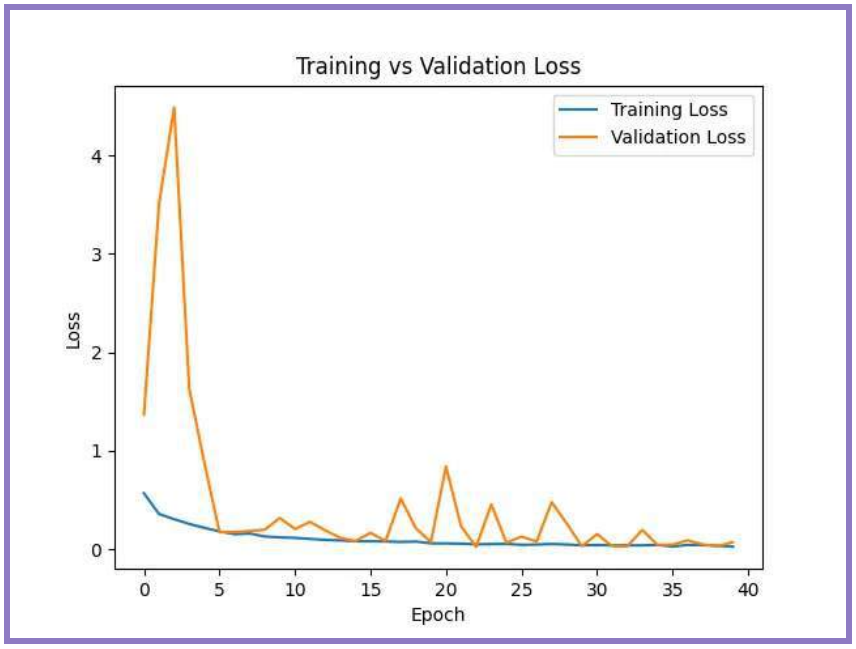
Confusion Matrix



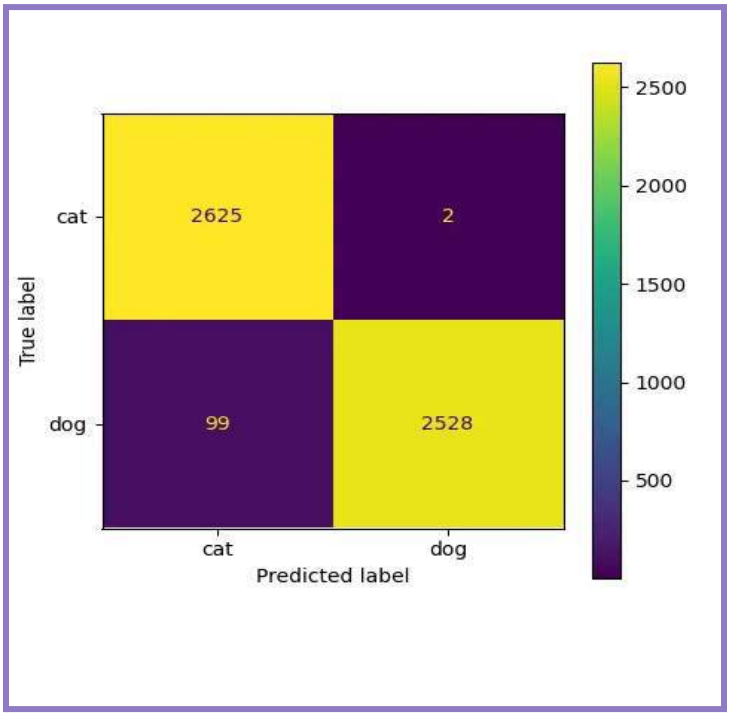
Dataset 2: Cats vs Dogs (No. of Files: 5254)
Training Statistics

- Training Configuration: 5 runs × 40 epochs
- Training Time: 2260.2 seconds

Performance : Accuracy: 97.91%, Loss: 0.07



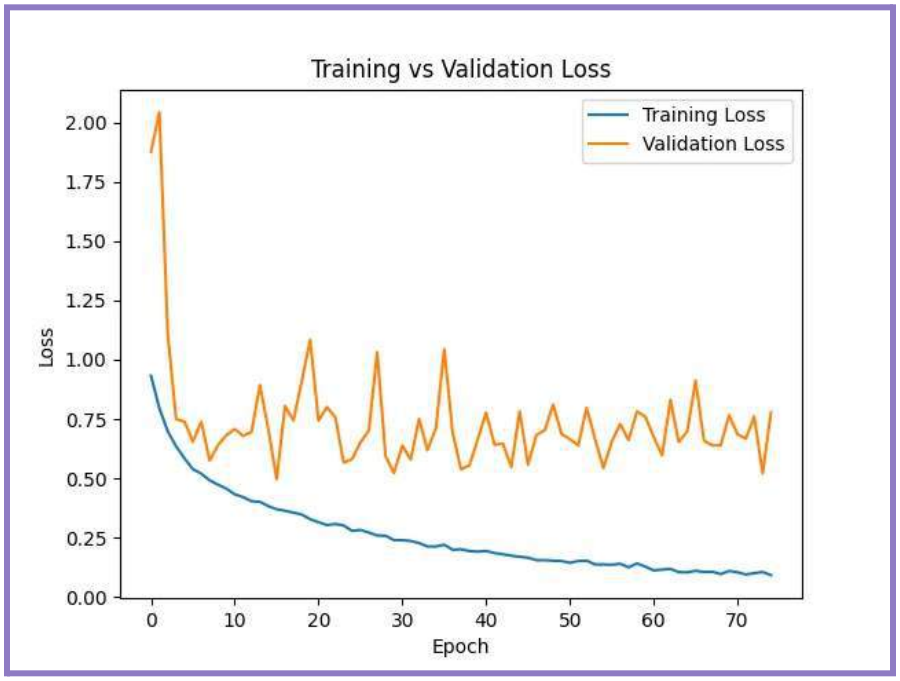
Confusion Matrix



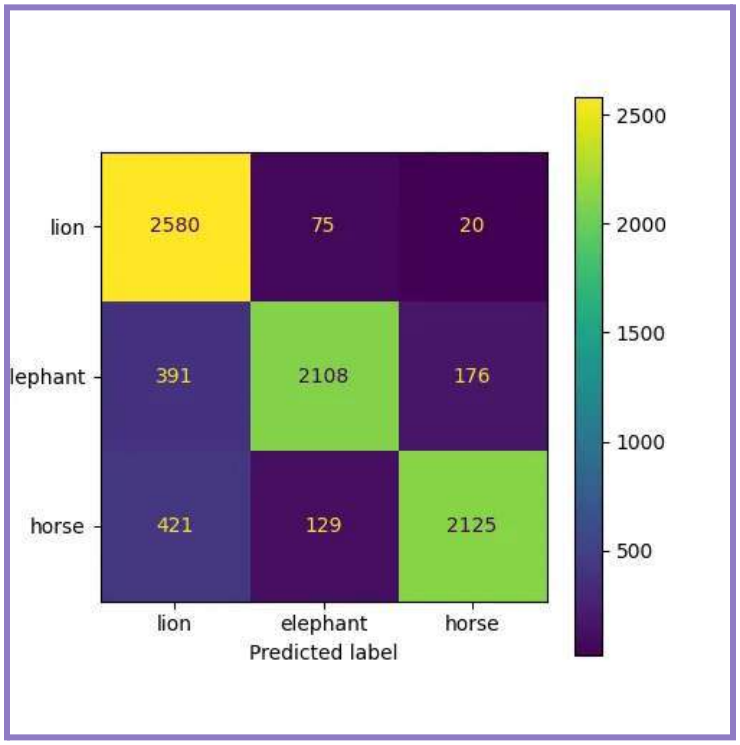
Dataset 3: Lion vs Horse vs Elephant (No. of Files: 8025)
Training Statistics :

- Training Configuration: 5 runs × 50 epochs
- Training Time: 5625.4 seconds

Performance : Accuracy: 79.25%, Loss: 0.78



Confusion Matrix

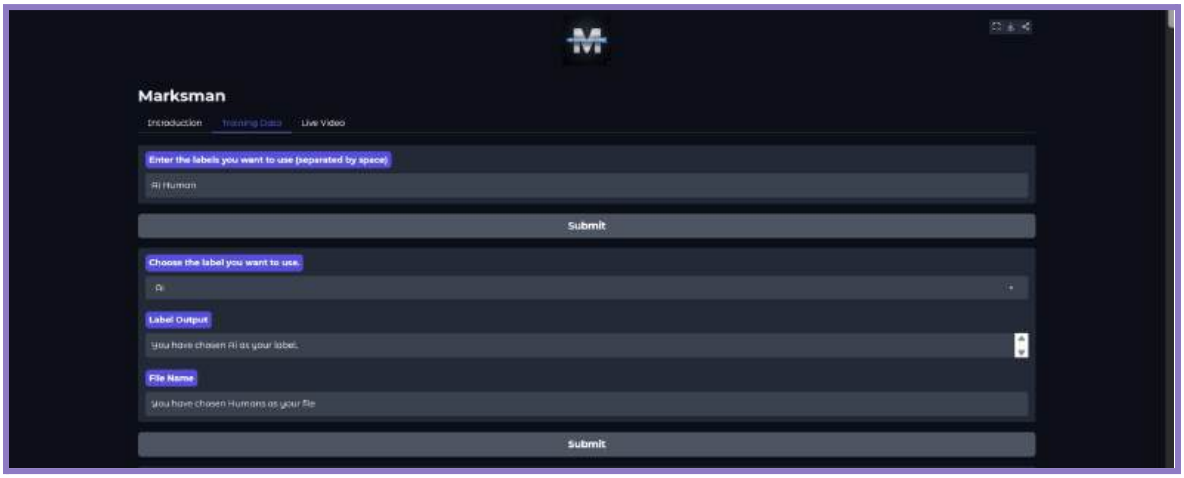


Based on the results above, we can see the model has an average accuracy above 90%, with the confusion matrices further confirming high correct classification rates across multiple

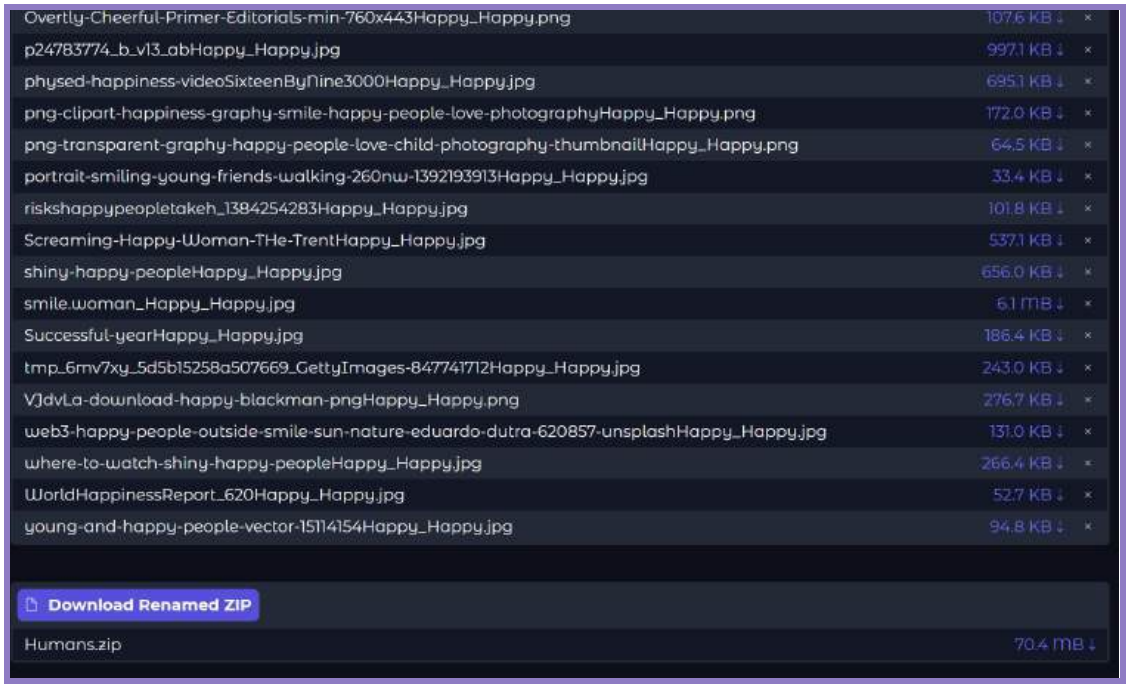
datasets. In addition, the cross-validation charts show strong model fitting, with only small differences between the training and validation loss. This indicates that the model is not overfitting and is able to generalize well to new, unseen data. Not only is the model achieving high accuracy, but it also maintains relatively low loss values, which means its predictions are both confident and precise.

Furthermore, the model performs consistently across datasets of different sizes and complexity. For example, it achieved over 97% accuracy on the Cats vs. Dogs dataset with more than 5,000 images, demonstrating scalability to larger datasets. Even in more complex multi-class scenarios such as Lion vs. Horse vs. Elephant, the model was still able to learn meaningful patterns despite increased difficulty and visual similarity between classes. These results demonstrate that the architecture is effective, reliable, and capable of supporting real-world image recognition tasks. Overall, this performance validates the model's ability to serve as a practical foundation for customizable object detection applications such as The Marksman.

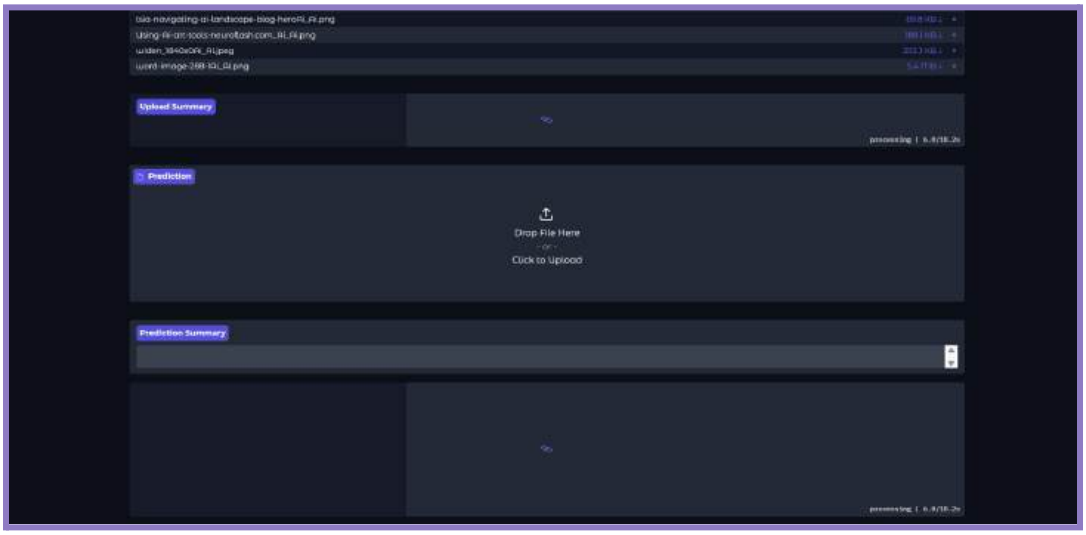
Application Images



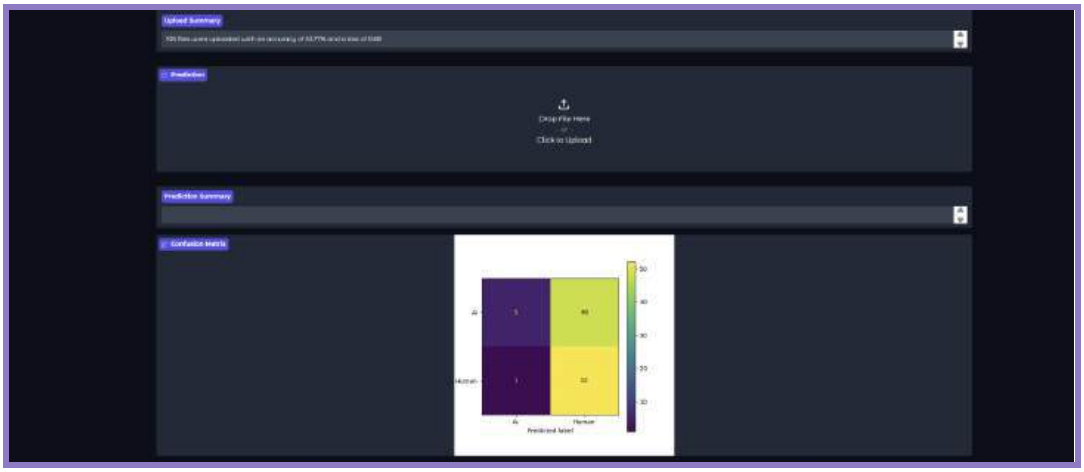
Setting up the model or data



Downloading New Renamed File



Processing Data/Uploading Images



Final Output of Training

Conclusion

The Marksman demonstrates that a customizable, AI-powered video recognition system can be effectively developed using a convolutional neural network and deployed through an accessible web application. The CNN architecture successfully learned visual patterns from user-provided datasets, achieving high accuracy across multiple classification tasks, including over 97% accuracy on larger datasets and strong performance even in more complex multi-class scenarios. The small gap between training and validation loss indicates strong generalization and minimal overfitting, supported by the use of data augmentation, Batch Normalization, and Dropout. By integrating the trained model into a Gradio-based interface with live webcam prediction, the system shows that advanced deep learning technology can be made practical and affordable for everyday users. Overall, the results validate that The Marksman provides a scalable, reliable, and customizable solution for real-time object detection and long-term monitoring applications.

Works Cites

Practical Deep Learning for Coders | Practical Deep Learning for Coders. (n.d.). Course.fast.ai. https://course.fast.ai/

Team, G. (2025). Blocks And Event Listeners. Gradio.app. https://www.gradio.app/guides/blocks-and-event-listeners

TensorFlow. (2019). Convolutional Neural Network (CNN) | TensorFlow Core. TensorFlow. https://www.tensorflow.org/tutorials/images/cnn

(2026). Huggingface.co. https://huggingface.co/docs/hub/spaces-sdks-gradio

Image recognition software, ML image analysis, and video analysis – Amazon Rekognition pricing – AWS. (2026, March 9). Amazon Web Services, Inc. https://aws.amazon.com/rekognition/pricing/?p=pm&c=ml&pd=rkg&z=4&refid=7c4209cd-a0c8-4a7d-9c3a-cb571e408c0c

The Effect of Low-Dose, Repeated Insect Repellent Exposure on Thermal Bleaching in Porites *cylindrica* By Issalyne Ladha

Abstract

Coral reefs are very diverse and productive ecosystems that are rapidly deteriorating due to chemical pollution and climate change. Recent studies have identified sunscreen chemicals as a very potent cause of coral bleaching and toxicity. While awareness of sunscreen's effects on coral reefs and other marine organisms has helped create coral-safe sunscreen products, the insecticides widely used along shorelines have not been well studied for their impact on coral health and aquatic life, nor is there a widely accepted, reef-safe substitute. This study examines the effects of common insecticides used along shorelines on the viability of coral fragments, particularly their symbiont populations, thereby evaluating their potential threat, comparable to that posed by sunscreen chemicals. Coral fragments were treated with different concentrations of a commercial insecticide. A symbiont population count was determined over two weeks using a hemocytometer, and coral condition was assessed visually; results were recorded. Some of the study samples were heat-treated to evaluate the effects of high tropical temperatures and global warming on insecticide-exposed coral. The results were analyzed using Analysis of Variance (ANOVA). The corals that received higher concentrations of the insecticide developed higher symbiont populations under stress than other samples and also showed the most visible modifications, with bleached and insecticide-exposed samples expelling 52 symbionts compared to 21 in insecticide-exposed samples. This study suggests that insecticides could pose a new threat to the chemical integrity of coral reefs and other ecosystems, highlighting the need for reef-safe substitutes.

Introduction

One of the pressing ecological issues currently threatening life on Earth relates to coral bleaching. At a fundamental level, it stems from an imbalance in the endosymbiotic relationship between corals and their zooxanthellae. The algae produce most of the energy necessary for survival through photosynthesis and, in turn, receive shelter and light from corals. As soon as stress occurs, perhaps due to global warming and pollution, corals expel these algae. Consequently, they lose color and vitality, a process known as bleaching. It leads to several effects, including slower growth, disease, reproductive failures, and even coral bleaching-induced extinction. Otherwise, it leads to an endosymbiotic collapse (Downs, Schroeder, and Brown, 2016).

Despite their delicacy and vulnerability, coral reefs are among the most vital ecosystems on Earth. Although they cover fewer than 1% of the total ocean surface, they are “biodiversity hotspots” that provide habitat, shelter, and nurseries for approximately a quarter of all marine species (Fabricius, Okaji, & De’ath, 2014). This makes them absolutely vital to sustaining global marine productivity and diversity. Humans also derive enormous benefits from these ecosystems. Reefs support fisheries that feed millions of people worldwide, drive multi-billion-dollar tourism

economies, and provide natural buffers against storm surge and erosion (Fabricius et al., 2014; Tarrant & Golbuu, 2023). Their loss would be an ecological catastrophe and an economic and societal disaster, specifically for coastal communities worldwide.

Although the impact of warmer oceans as a stressor inducing bleaching events is well understood, emerging literature highlights additional stressors, including chemical contaminants. These stressors are ubiquitous in today's oceans and primarily enter the oceans through land runoff and shore-based activities. Although these stressors are sublethal, they still affect coral physiology and functions. Examples include inhibition of photosynthesis and metabolism and acceleration of slow growth rates, making corals more susceptible to stress from high temperatures. According to various researchers, these stressors include sunscreens, insect repellents, pharmaceuticals, and pesticides.

Chemical Effects

DEET (N, N-diethyl-meta-toluamide), an insect-repelling compound commonly found in insect repellents, and oxybenzone (benzophenone-3), a UV protector previously common in consumer-grade sunscreens, are among the most widely documented pollutants. DEET primarily enters marine and freshwater ecosystems via swimmers' wash-off, sewage effluent, and runoff from regions with high rates of insect repellent use. Noting a marked difference between DEET and most sunscreen compounds, DEET, as relatively water-soluble and persistent, can migrate more systematically within coastal ecosystems and bioaccumulate within seawater, sediments, and biota (Waller & Mahon, 2020). DEET has been shown to compromise photosynthesis, metabolic function, and physiological performance, particularly under intense heat. At higher concentrations, DEET has been associated with biocoloration, reduced calcification rates, and heightened sensitivities to co-occurring contaminants, implying a 'synergistic' effect with warming oceans that reduces coral resilience.

Oxybenzone remains a documented coastal pollutant in some areas, primarily due to runoff from swimming pools and sewage treatment plants. Although present at low concentrations, it still induces oxidative stress, DNA damage, and developmental aberrations in coral larvae (Downs et al., 2016; Stien et al., 2024). Moreover, it could stimulate photosensitization reactions, thereby exacerbating photosensitivity-induced cell injury. It has been observed that its distribution is decreasing in regions with regulated sunscreen use.

These substances also have detrimental effects on humans. Both DEET and oxybenzone are absorbed through the skin and have been detected in urine, blood, and breast milk. Oxybenzone has been examined for its endocrine-disrupting properties (Krause et al., 2012), and excessive or prolonged DEET exposure has been associated with neurological problems. However, it is considered safe for use at typical concentrations for personal and cosmetic services. Various research reports indicate that DEET may cause skin reactions and, when combined with other ingredients in personal care products, may stimulate systemic absorption.

Effects of Insecticides and Heating Stressors

To date, mitigation actions have brought forth alternatives. Regarding insect repellents,

picaridin and oil of lemon eucalyptus have emerged as promising, more environmentally acceptable alternatives that degrade more readily and exhibit lower aquatic toxicity than DEET. Regarding sunblocks, mineral sunblocks based on non-nano zinc oxide and titanium dioxide offer reef-safe, physical UVA/UVB protection and are less reactive with water. By promoting changes and adopting preventive measures such as wearing UVB-protective clothing, staying indoors during peak sun hours, and selecting eco-labeled products, the release of chemicals into marine ecosystems can be avoided or minimized (Stien et al., 2024).

Notably, coral stress responses are usually not caused by a single factor. Complex interactions occur. It has been found that simultaneous stress from increased temperatures and pollutant exposure amplifies the stress response beyond that caused by either component alone (Le Nohaïc et al., 2017; Stein et al., 2024). Notably, these findings have profound implications for corals living in coastal regions, given the prevalence and severity of thermal stress. As suggested by research on oxybenzone and other components at previously determined trace levels, these components have been found to contribute to decreases in coral cover (Tarrant & Golbuu, 2023). Aside from these known stressors, emerging threats are further challenging reef health. Microplastics, widely dispersed throughout the oceans, affect feeding efficiency, energy balance, and coral physiological functions (Ziduluo et al., 2025). They can also transport other pollutants.

Despite decades of extensive research, knowledge gaps persist. The majority of research efforts have focused on specific stress factors, and surprisingly few studies have examined low-dose contaminant exposures and their interactions with thermal stress. This is of remarkable importance because corals are rarely exposed to high concentrations of a single contaminant in reef environments. Corals are repeatedly exposed to runoff, sewage, and other human impacts in coastal regions. The present research aims to fill this research gap by examining the effects of repeated low-dose exposure to a commercial insect repellent on the bleaching response of *Porites cylindrica*. To examine the effects of various environmental and climatic conditions, experiments were conducted at different temperatures. The health status of corals will be monitored by observing changes in pigmentation and the degree of algal expulsion, both of which are considered vital factors in bleaching. By doing so, it will be possible to understand how rising temperatures might be exacerbated by chemical pollution.

Research Questions

RQ 1: How does repeated low-dose exposure to insect repellent affect bleaching severity in *Porites cylindrica*?

RQ 2: Does exposure to insect repellent increase the expulsion of symbiotic algae (symbionts) under thermal stress?

RQ 3: What is the combined effect of insect repellent and heat on coral health indicators such as color and symbiont density?

Hypothesis

H1: Repeated low-dose exposure to insect repellent will increase the severity of coral bleaching

and symbiont loss in *Porites cylindrica* under elevated temperature conditions
H2: Combined chemical and thermal stress will have a synergistic adverse effect on coral health compared with the impact of either stressor alone.
H3: DEET is the specific chemical in insecticides that damages coral tissue and is also harmful to other species.

Methodology

To explore the effects of insecticides on coral health, the student prepared dilutions, applied treatments, cleaned plates, photographed the corals, and counted symbionts under a microscope. The student also did all data collection and analysis. The mentor operated and supervised the heating equipment, ensuring proper temperature control and safety. In addition, the mentor provided coral samples.

Experimental Design

Two 12-well plates containing small fragments of *Porites cylindrica* were prepared. One plate was subjected to elevated temperature to simulate heat stress, while the other plate remained at ambient temperature. Coral fragments were assigned to treatments consisting of three dilutions of a commercial insect repellent (1:500, 1:1000, 1:5000) and a control group with no repellent. Each treatment and control was applied to three replicate wells per plate.

Procedure

Coral fragments with minimal discoloration were selected to ensure initial health and consistency. Plates were gently cleaned using a pipette and brush to remove detritus without damaging the coral tissue. Baseline photographs were taken to document initial coloration. Treatments were applied by removing water from each well and adding the corresponding insect-repellent dilution. Coral fragments were exposed to 30 minutes of absorption, rinsed, and returned to their respective tanks. Treatments were repeated every three days for a total of four rounds.

The heat treatment was applied using a heating device operated by a mentor. Symbiont expulsion was counted under a microscope (hemocytometer), and changes in coral coloration were quantified using a grayscale color scale (0 = black, 256 = white).

Data Analysis

Descriptive statistics summarize means and variability for each treatment group. A standard deviation graph was created based on the symbiont counts graph. ImageJ was used to analyze color using a color scale, focusing on red, blue, and green colors. Data were visualized using bar graphs and line plots to illustrate treatment effects over time.

Safety Considerations

In a well-ventilated laboratory, all chemical handling was performed with nitrile gloves and protective eyewear. Coral fragments were handled carefully to avoid damage, and the

heating device was operated only by a mentor. Standard laboratory safety protocols were followed, including the quick cleanup of spills and the prohibition of food and drink in the lab.

Results

Coral fragments exposed to repeated low-dose insect repellent showed clear, concentration-dependent increases in bleaching intensity and algal symbiont expulsion. These effects were most shown under elevated temperature conditions. The following data summarize changes in pigmentation (grayscale color value) and the average number of symbiont expulsions per 100 µL sample for each treatment.

As shown in Table 1, higher concentrations of insect repellent (1:500) resulted in the lightness coloration (i.e., higher grayscale values), indicating symbiont loss.It is also important to note that higher grayscale values indicate greater bleaching (lighter coloration). Symbiont expulsion increased with both chemical concentration and temperature. In contrast, the control and 1:5000 dilution showed minimal bleaching, especially under ambient conditions. Symbiont expulsion counts followed the same pattern as shown in Figure 1, showing that both higher concentration and elevated temperature caused increased expulsion of algal symbionts, indicating stress-induced bleaching.

Treatment (Dilution)	Temperature Condition	Mean Grayscale Value (±SD)	Mean Sybiont Count (±SD)
Control	Ambient	45 ± 3	12 ± 2
1:5000	Ambient	53 ± 4	13 ± 2
1:1000	Ambient	68 ± 5	27 ± 3
1:500	Ambient	82 ± 6	35 ± 4
Control	Elevated	71 ± 5	21 ± 3
1:5000	Elevated	95 ± 4	40 ± 4
1:1000	Elevated	124 ± 7	47 ± 5
1:500	Elevated	148 ± 5	52 ± 6

Table 1: Mean grayscale color values (0–256) and average symbiont expulsion counts per 100 µL sample for Porites *cylindrica* across treatments and temperatures. (Table created by student)

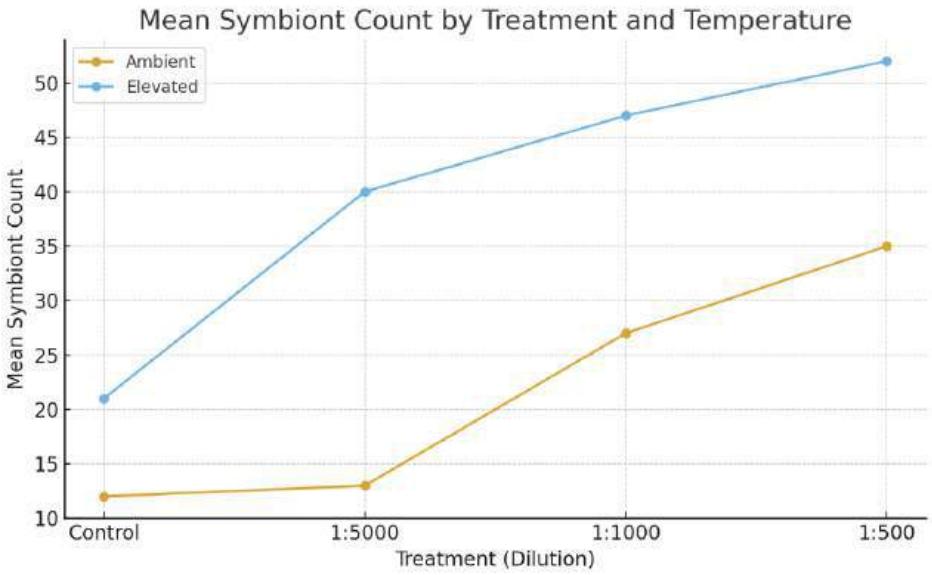
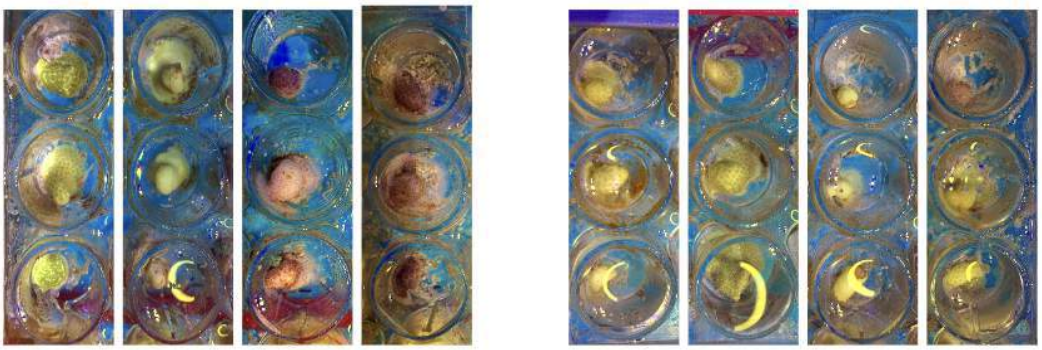


Fig 1: Mean symbiont expulsion per treatment under ambient and elevated temperature conditions. (Figure created by student)

Mean symbiont counts increased with both higher chemical concentration and elevated temperature. Under ambient temperature, symbiont counts ranged from 12 ± 2 in the control to 35 ± 4 in the 1:500 treatment. In contrast, elevated temperature produced consistently higher counts across treatments, ranging from 21 ± 3 in the control to 52 ± 6 at the highest concentration.

This pattern shows an apparent combined effect of treatment and temperature, with elevated temperature amplifying the increase in symbiont counts already observed across dilutions. Although the elevated groups appear higher visually, a t-test comparing ambient versus elevated temperature did not reach statistical significance ($t = –2.07$, $p = 0.0834$), suggesting the observed differences may represent a trend rather than a statistically significant effect.

Elevated Temperature: Ambient Temperature:



Control 1:5000 1:1000 1:500 Control 1:5000 1:1000 1:500

Image 1: Coloration variation showing damage from thermal and chemical stress. (Images taken by the student)

These images display the coloration of each dilution for both elevated and ambient temperatures. When viewed from left to right, the control shows no whitening or darkening, indicating no bleaching or death of the fragments. In the elevated-temperature group, there is rapid darkening, indicating death, especially in the corals from the 1:1000 and 1:5000 dilutions, which contain the highest insecticide concentrations. Looking at the ambient temperature group, the fragments appear whitened, indicating slight bleaching.

The color variation shows that when the insect repellent is mixed with thermal bleaching, representing rising ocean temperatures, the damage to the coral becomes visibly worse, even past the point of bleaching. Even at low doses, a synergistic effect occurs between the insecticide's chemicals and the elevated water temperature.

Discussion

The result of this research is that repeated low doses of a commercial insect repellent significantly increased coral bleaching in Porites cylindrica, and this effect accelerated with increasing temperature. Coral samples exposed to higher insect repellent doses exhibited the greatest loss of color and expulsion of symbionts, thereby reinforcing the dose-dependent relationship.

Dose Response

These findings confirm hypothesis 1 (H1), which predicted that repeated low doses of an insect repellent would stress corals and enhance bleaching levels and loss of symbiotic algae at higher temperatures. The extent and nature of coral bleaching, particularly within the 1:500 solution, clearly show that even very low doses of common consumer chemicals can increase physiological stress in corals exposed for prolonged periods (Cunning et al., 2017).

Chemical and Thermal Stress

The data clearly validate hypothesis 2 (H2), which suggested that the combination of chemical and thermal stress would have a synergistic damaging effect on corals compared with stress alone, in particular, under heat stress, loss of color and expulsion of symbionts escalated significantly for all concentrations, indicating that heating intensified the impact of chemicals. The greatest degree of bleaching was observed at a 1:500 dilution under heat stress, with symbiont expulsion exceeding 3.5-fold that measured under unheated control conditions. The presence of a synergistic relationship indicates that pollutants can increase the thermal stress required for bleaching, as corals exposed to chemicals may start bleaching at lower temperatures than unexposed corals. Similar synergistic interactions have been previously documented among various pollutants, such as oxybenzone and DEET and rising sea temperatures (Le Nohaïc et al., 2017; Stein et al., 2024; Sturvan et al., 2024).

Chemical interpretation

Finally, hypothesis 3 (H3) suggested that DEET harms coral tissues and could affect various species in marine ecosystems. Although it cannot be ascertained from the current

research that DEET alone contributed as an independent variable within a specific insect repellent product, it can be determined that the bleaching effects seen within the research correspond with acknowledged DEET toxicity, such as DNA damage, oxidative stress, and coral-algal symbiosis disruption (Downs et al., 2016). Although it cannot be ascertained at this stage whether DEET significantly contributes to insect repellent efficacy per se, there can be no denying that personal products containing DEET and similar compounds can have potentially harmful effects on an ecological scale in a marine setting.

These results together make it clear that LD exposures can be as damaging as HD exposures, and that these effects can be exacerbated with a warming background. The similarity in trends between the color and symbiont datasets indicates that these contaminants can affect coral resilience, making it more susceptible to challenges such as rapid temperature increases. Coral bleaching is never caused by a single stressor, but rather by a combination of various human-induced stresses (Good et al., 2021).

Limitations and Challenges

Although the data demonstrated tendencies in coral status and insect-repellent use, these are based on small sample sizes. Only two plates containing coral fragments were used, and differential effects among individuals may have occurred during the bleaching tests (Dobson et al., 2023). It was sometimes hard to maintain a constant temperature with the heating method. Given the experiment's short duration of several weeks, testing for longer-term effects, such as recovery and delayed bleaching, would have been appropriate. It was also a disadvantage that color testing relied on picture-interpretable data and grayscale values, as some data might have been more visible than others. Because of these limitations, developing a data analysis based on color differences among the coral would have been difficult, and creating a graph based on these attributes would not have been possible. Despite these limitations, it can be noted that the experimental setup effectively controlled for and isolated the main factors, which were chemical concentration and temperature, with authentic outputs.

There were also some minor setbacks and changes incorporated into the experiment. For instance, some small pieces of coral broke off during rinsing and had to be replaced. This led to some changes in how samples were processed for subsequent experiments to avoid stressing the coral. A special glue called ecotech had to be used for attaching these small bits to the wells.

Future Research

To gain insight into the effects of prolonged exposure, the scope will be extended to include additional species and longer exposure durations to assess whether any recovery or adaptation occurs after pollutant challenge. A more realistic set of results might be achieved by examining multiple pollutant types or by combining various chemicals. It would be helpful to obtain more information on which biochemical changes in coral physiology are occurring as a result of repellents, as measured, for example, by photosynthetic efficiency. Field experiments would also help determine whether a bleaching response similar to that observed in a laboratory setting occurs.

Interpretation and Implications

The result supports previous studies on the importance of chemical pollution for resilience and the white bleaching phenomenon. The results indicated that controlling sources of chemical pollution, for example, everyday products containing insect repellents and sunscreen, would be a focus for reducing stress on coral reefs. Given its direct association with human activities, it is among the most feasible methods for conserving coral species (Wilkinson, 2011). It also confirms that multiple factors cause white bleaching in coral species.

Conclusion

This research aimed to understand the effects of low doses of a commercial insect repellent on thermal bleaching of Porites cylindrica, an essential coral species found in coral reef ecosystems. Findings clearly showed that low chemical doses can substantially accelerate bleaching, and this effect becomes more pronounced at higher temperatures. Higher concentrations of the repellent triggered greater pigment reduction and greater algal removal from coral tissues, confirming that stress-inducing chemicals accelerate the physiological stress caused by high temperatures. These observations are consistent with previous research on specific contaminants, such as oxybenzone and DEET, which are typically associated with high concentrations due to their widespread use in sunscreens and insect repellents.

By recreating a natural environment, this experiment connects a crucial research gap in coral studies by investigating how corals respond to repeated daily pollution rather than to single events. These findings imply that even low levels of pollution can cause coral bleaching over time, thereby increasing the susceptibility of coral ecosystems to rising temperatures driven by climate change. Understanding that it is not just temperature but multiple human-introduced stressors that contribute to coral deterioration lays a foundation for local and global efforts to control the use and distribution of these harmful chemicals and to examine how reef-safe alternatives interact with natural environments. Another area of research could be applying a similar concept to reef-safe sunscreen and developing reef-safe insect repellents. The product would employ alternatives that exclude oxybenzone, DEET, and other harmful pollutants, serving as a dual protector against insect bites for humans and against adverse effects on coral and other marine life.

Another innovation in this experiment was the use of wells filled with coral fragments. The technique of using wells with coral pieces has never been tested before. However, it works very well because it allows for isolation, so that different wells have different concentrations or solutions. The wells also carry the symbionts for every piece of coral. It is thus easy to calculate the number of symbionts associated with a particular piece of coral. Before that, the symbionts would have been drifting in a tank, making it challenging to identify which symbionts belonged to a specific piece of coral. Effectively, this research highlights, ever more dramatically, the importance of reef preservation as an ecological and economic life-support system. By better understanding and framing the role that common contaminants play in amplifying the effects of stress caused by global warming, there is an ever clearer and more hopeful road ahead toward finding more sustainable practices for humanity.

Works Cited

Downs, C. A., et al. (2016). Toxicological effects of the sunscreen UV filter, oxybenzone (benzophenone‐3), on coral planulae and cultured primary cells, and its environmental contamination in Hawaii and the U.S. Virgin Islands. Archives of Environmental Contamination and Toxicology, 70(2), 265–288. https://doi.org/10.1007/s00244-015-0227-7

Fabricius, K. E., et al. (2014). Ecological effects of chemical pollution on coral reefs: Exposure and ecological effects of personal care products and pharmaceuticals. Marine Pollution Bulletin, 85(1), 27–40. https://doi.org/10.1016/j.marpolbul.2014.04.028

Le Nohaïc, M., et al. (2017). Coral bleaching response to thermal stress under varying chemical exposure. Scientific Reports, 7(1), 13221. https://doi.org/10.1038/s41598-017-13114-9

Navon, N., et al. (2024). Detection of common pharmaceuticals in reef‐building corals. Environmental Pollution, 347, 123780. https://doi.org/10.1016/j.envpol.2024.123780

Negri, A. P., & Jones, R. J. (2015). Chemical pollution and coral reef bleaching: Sub‐lethal effects of contaminants on coral photosynthesis. Coral Reefs, 34(4), 991–1002. https://doi.org/10.1007/s00338-015-1307-6

Rao, J. V., & Rajeswari, S. (2019). Insecticides and a fungicide affect multiple stages of coral life. Environmental Pollution, 252, 258–267. https://doi.org/10.1016/j.envpol.2019.05.02

Stien, D., et al. (2024). Sunscreens: Potential hazards to environmental and human health. Frontiers in Marine Science, 10, 1471574. https://doi.org/10.3389/fmars.2024.1471574

Tarrant, A. M., & Golbuu, Y. (2023). Fine‐scale geographic risk assessment of oxybenzone in coral reef environments. Environmental Toxicology and Chemistry, 42(3), 657–669. https://doi.org/10.1002/etc.5607

Waller, R. G., & Mahon, A. (2020). Bacillus insecticides are not acutely harmful to corals and sponges. Journal of Applied Microbiology, 129(5), 1234–1244. https://doi.org/10.1111/jam.1467

Ziduluo, X., et al. (2025). Tracking microplastics in live reefs‐building corals with microscopic laser particles. Marine Pollution Bulletin, 200, 115634. https://doi.org/10.1016/j.marpolbul.2025.115634

Genetic and Biological Basis of Lean Type 2 Diabetes in South Asian Populations: A Narrative Review By Adithi Vattem

Abstract

Type 2 diabetes is highly prevalent in South Asian populations, who also develop the condition much younger and at lower BMIs than other groups. Additionally, these statistics challenge traditional obesity-based models of Type 2 Diabetes, which are primarily modeled off of European cohorts alone. This review examines the genetic and biological factors underlying lean Type 2 diabetes (T2D). First, indicated by evidence, lean and obese T2D share common genetic risk variants, however these variants have a stronger effect on lean individuals, which reflects a reduced metabolic tolerance rather than a distinct genetic disease. Second, a smaller group of lean patients may carry rare high-impact mutations in β-cell genes that have gone undetected by the standard GWAS. Third, slight genetic variation in lipodystrophy-related pathways may impair adipose tissue function, as a result leading to ectopic fat and insulin resistance at a normal or lower body weight. Finally, these findings seem inconsistent with the thrifty genotype hypothesis, since the hypothesis predicts obesity-driven insulin resistance rather than the insulin deficiency pattern which was observed in lean Indians. Overall, lean T2D in South Asians reflects an overall mix of genetic risk, reduced metabolic tolerance, impaired fat storage, and rare mutations in some situations, ultimately highlighting the need for population-specific studies to understand and treat T2D.

Keywords

Type 2 diabetes; South Asians; lean diabetes; insulin deficiency; β-cell function; lipodystrophy; polygenic risk; population-specific genetics

Introduction

Among all ethnic groups, South Asian populations face one of the highest burdens of Type 2 diabetes mellitus (T2D), developing the disease at younger ages and at significantly lower body mass indices (BMI) (Mohan). Unlike the classic obesity-driven model of T2D, a substantial proportion of South Asian individuals develop diabetes despite being lean or only mildly overweight. Studies show that 24–39% of Indian patients with diabetes have a normal BMI, compared with approximately 9% in European populations (Siddiqui et al.). These individuals frequently present with earlier disease onset, reduced insulin secretion, and more rapid progression to insulin dependence (Mohan; Salvatore et al.), suggesting a distinct biological basis for disease development. This review focuses primarily on evidence from Indian cohorts, with reference to broader South Asian data where applicable.

In addition to genetic susceptibility, impaired adipose tissue function and reduced fat storage capacity may contribute to metabolic dysfunction in lean individuals (Stefan et al.; Misra and Khurana), even in the absence of excess body fat. Together, these observations suggest that lean Type 2 diabetes in South Asians cannot be explained by obesity alone, and raise the

question of whether lean individuals with T2D have a distinct genetic and biological basis for the disease.

This review examines the genetic and biological basis of lean T2D in Indian populations across four areas: whether genetic risk factors are shared with obese T2D, the role of rare high-impact mutations, the contribution of adipose tissue dysfunction through lipodystrophy-related pathways, and whether population-specific evolutionary models such as the thrifty genotype explain the observed pattern.

Shared Genetic Architecture in Lean and Obese T2D

The first possible explanation for shared genetic architecture between lean and obese T2D is that lean individuals experience a greater obesity-driven metabolic stress, despite having normal BMI, which leads to earlier diagnosis of diabetes. The available evidence does not support this interpretation, since it is observed that lean T2D is far more common in Indians than Europeans, despite the absence of obesity (Siddiqui et al.). Lean Indian patients also exhibited reduced stimulated C-peptide levels (small proteins which are produced along with insulin after eating) (Prasad et al.). This indicated severe insulin deficiency rather than obesity-driven insulin resistance. Had metabolic stress been the primary driver, diabetes would be expected to occur mainly in individuals with both a high adiposity and greater insulin resistance, which was not observed in the studies done. Moreover, analysis of South Asian populations show that individuals with high genetic risk develop diabetes about eight years earlier, along with a BMI 3 kg/m^2 lower than their counterparts, suggesting that diabetes can begin without obesity because the body's ability to compensate is already impaired (Hodgson et al.). These findings would ultimately argue against the idea that hidden adiposity or excess metabolic load is a contributor to lean Type 2 Diabetes.

A second possibility is that lean Indian T2D represents a genetically distinct form of diabetes driven by unique risk variants. However, genetic studies do not support the existence of a separate disease entity. A large BMI-stratified genome-wide association study demonstrated that known T2D risk alleles are largely shared between lean and obese individuals, with no evidence of completely distinct genetic drivers (Perry et al.). Instead, the effects of these shared risk variants were significantly stronger in lean individuals, indicating that obesity can partially mask underlying genetic susceptibility. Similarly, South Asian–specific GWAS have identified diabetes-associated loci that overlap with established T2D pathways, particularly those related to insulin secretion and β-cell function, rather than identifying an entirely novel genetic disease (Kooner et al.). These findings suggest that lean Indian T2D shares a common genetic foundation with obesity-associated T2D, differing primarily in the magnitude and biological weighting of genetic risk rather than in gene identity.

Since both increased metabolic stress and genetic distinctness have been excluded, the remaining explanation is that lean Indian individuals have a reduced physiological tolerance to shared diabetes risk. In this model, genetic variants disproportionately impair systems responsible for maintaining metabolic stability, such as insulin secretion and adipose tissue function. Pathway-partitioned polygenic analyses in South Asian populations demonstrate that

genetic risk is enriched in β-cell dysfunction and unhealthy fat-distribution pathways, which are associated with earlier disease onset, faster progression, and lower BMI at diagnosis (Hodgson et al.). These pathways directly influence the body's ability to compensate for modest insulin resistance or lipid spillover. Supporting this interpretation, transcriptomic analyses of adipose tissue in Indian T2D patients reveal dysregulation of lipodystrophy-related genes, indicating impaired fat storage capacity despite normal or low overall fat mass (Saxena et al.). When adipose tissue cannot safely buffer excess lipids, ectopic fat deposition occurs at lower levels of adiposity, accelerating insulin resistance and β-cell stress.

Rather than arising from greater metabolic stress or a distinct genetic disease, lean Indian T2D reflects a reduced metabolic tolerance to shared genetic risk. Inherited defects in insulin secretion and fat storage reduce physiological buffering capacity, allowing diabetes to manifest at normal body weight. This framework reconciles shared genetic architecture with earlier disease onset and lower BMI in Indian populations and provides a coherent explanation for why diabetes develops in the absence of obesity. However, this analysis has focused exclusively on common variants; the genetic architecture of lean T2D may also include rarer mutations with larger individual effects. This interpretation is further supported by a large comparative analysis of Indian and European cohorts, which independently concluded that lean T2D in Indian populations is driven by impaired insulin secretion and higher β-cell genetic risk burden rather than by distinct genetic variants or excess adiposity (Hodgson et al.).

Rare Mutations in Lean T2D

The preceding analysis focused on common variants detectable by GWAS, however a subset of lean patients may carry rarer, higher-impact mutations which operate through a distinct mechanism. Recent sequencing studies suggest that a subset of lean diabetes patients may carry rare or damaging genetic mutations that are not detected in standard genome-wide association studies. While GWAS are highly effective at identifying common variants that contribute small increases in disease risk, they are not designed to detect rare coding mutations that may have much larger biological effects. Large-scale exome sequencing analyses have shown that rare loss-of-function or missense variants in genes involved in pancreatic β-cell function can substantially increase diabetes risk, often with effects far stronger than typical GWAS variants (Flannick et al.). These damaging mutations frequently affect pathways related to insulin secretion, β-cell survival, and glucose sensing, suggesting that insulin deficiency rather than obesity-driven insulin resistance may be a primary driver of disease in certain individuals. Because these variants are extremely rare within populations, they often remain undetected in large association studies that focus primarily on common genetic variation.

Evidence from clinical sequencing studies further supports the presence of rare high-impact mutations in individuals diagnosed with type 2 diabetes. Genetic screening of patients with T2D has identified pathogenic mutations in genes associated with monogenic diabetes (MODY), particularly among individuals who are lean and diagnosed at younger ages (Kwak et al.). Many of these mutation carriers had originally been classified as having

conventional type 2 diabetes despite carrying single-gene defects that directly impair insulin production. This suggests that misclassification may occur in clinical settings where genetic testing is not routinely performed, particularly in populations where lean diabetes is common.

Exome sequencing studies also demonstrate that lean or early-onset diabetes cases may carry a higher burden of damaging rare variants in genes involved in pancreatic β-cell biology (Flannick et al.). These variants include missense mutations and loss-of-function alleles that alter protein structure or disrupt gene function, many of which are not detectable through traditional GWAS approaches. The findings suggest that diabetes exists along a genetic continuum ranging from classical monogenic disorders to highly polygenic disease, with some lean diabetes cases occupying an intermediate position along this spectrum.

Adipose Tissue Dysfunction and Lipodystrophy-Related Pathways

Another potential explanation for lean Type 2 diabetes in Indian populations is the presence of subtle genetic variation in lipodystrophy-related pathways that alter how adipose tissue stores and distributes fat. Lipodystrophy refers to a group of metabolic disorders characterized by impaired fat storage, which leads to ectopic lipid accumulation in organs such as the liver and muscle and ultimately causes severe insulin resistance. Although classical lipodystrophy is typically caused by rare single-gene mutations, researchers have proposed that lean Type 2 diabetes in Indian populations may represent a milder, polygenic form of adipose tissue dysfunction (Saxena et al.; Stefan et al.).

Evidence supporting this hypothesis comes from transcriptomic analyses of adipose tissue in Indian diabetes patients. Studies examining peripheral subcutaneous adipose tissue identified differential expression of multiple genes previously associated with lipodystrophy syndromes. Among 138 known lipodystrophy-related genes analyzed, 16 were significantly dysregulated in individuals with Type 2 diabetes compared with healthy controls, and approximately 18% of these genes overlapped with genes already associated with diabetes-related metabolic pathways (Saxena et al.). Importantly, the expression of these genes correlated with metabolic traits such as glucose regulation and insulin sensitivity in diabetic individuals but not in healthy controls, suggesting these genes may play a role in the development of the disease. These findings suggest that adipose tissue dysfunction in lean diabetes may arise from subtle disruptions across many genes involved in fat storage and adipocyte biology rather than from a single high-impact mutation.

Genetic association studies further support the role of subtle metabolic risk variants in lean Indian diabetes. As previously noted, lean young-onset diabetes is substantially more common in Indian cohorts than European ones (Prasad et al.). Individuals with normal BMI in these Indian cohorts also exhibited significantly lower stimulated C-peptide levels (Siddiqui et al.), indicating impaired insulin secretion rather than obesity-driven insulin resistance. β-cell–specific polygenic risk scores, discussed further in the following section, also suggest that inherited defects in insulin production interact with impaired adipose storage capacity (Hodgson

et al.). These results suggest that inherited metabolic risk may interact with impaired adipose storage capacity, allowing diabetes to develop even in the absence of excess body fat.

Genome-wide association studies also identify variants that influence both adipose tissue biology and diabetes susceptibility, further supporting this model. Several loci associated with Type 2 diabetes risk are involved in pathways regulating fat distribution, adipocyte differentiation, and metabolic regulation rather than simply overall adiposity (Kooner et al.). These findings suggest that genetic risk in lean Indian diabetes may partly arise from variants that alter how adipose tissue develops and responds to metabolic stress.

Although this model shares the idea that population-specific genetic architecture contributes to diabetes risk with the thrifty genotype hypothesis, the two frameworks predict different metabolic outcomes. The thrifty genotype predicts enhanced energy storage and obesity-driven insulin resistance, meanwhile dysfunction in adipose tissue, which was observed in lean Indian cohorts, predicts impaired fat buffering, ectopic lipid deposition, and insulin deficiency.

Insulin Deficiency Versus the Thrifty Genotype Hypothesis

Scientists have suggested that the high rate of Type 2 diabetes in Indian populations may be related to the Thrifty Genotype Hypothesis, which proposes that some populations evolved genes that promote efficient energy storage during periods of famine (Neel). While this may have been advantageous historically, it could increase the risk of diabetes in modern environments.

However, recent evidence suggests that this explanation does not fully account for the pattern observed in Indian populations. If the thrifty genotype were the main cause, diabetes in Indians should look like obesity-driven insulin resistance from storing too much energy. Instead, lean Indian patients show reduced C-peptide responses. This points to insulin deficiency, which is the opposite of what the hypothesis predicts (Siddiqui et al.). Because the observed pattern is the reverse of the predicted pattern, the thrifty genotype does not explain diabetes in this population.

A genetic analysis by Yajnik and colleagues used a β-cell partitioned polygenic score and found that South Asians carry a higher burden of genetic variants associated with impaired insulin secretion. The effect of these variants was strongest in individuals with normal BMI and was linked to earlier diagnosis, suggesting that genetic risk related to β-cell dysfunction plays a particularly important role in lean individuals (Yajnik et al.).

Further support comes from the WellGen study, which analyzed 1,612 Indian patients with young-onset type 2 diabetes. Severe insulin-deficient diabetes (SIDD) was the most common subtype, accounting for 53% of patients versus 38% for the mild obesity-related subtype, and was characterized by early onset and reduced insulin production (Prasad et al.). The researchers also examined genetic risk scores for type 2 diabetes, BMI, lipids, and insulin-related traits, and found that these scores behaved differently in Indian populations compared to Europeans (Yajnik et al.). This suggests a population-specific genetic architecture, further supporting the idea that insulin deficiency is a key driver of diabetes in this group.

Genome-wide association studies in South Asian populations provide additional evidence. A large study involving over 50,000 individuals identified several novel type 2 diabetes risk loci, including GRB14, ST6GAL1, VPS26A, HMG20A, AP3S2, and HNF4A (Kooner et al.). Many of these genes are involved in β-cell function and insulin signaling rather than energy storage. These findings demonstrate that genetic risk for diabetes in South Asians includes variants that directly affect insulin production and regulation, and that such variants may not be captured in studies focused on European populations.

Conclusion

Lean Type 2 diabetes in Indian populations is best explained by a combination of shared genetic risk with amplified effects, rare high-impact mutations in some cases, and impaired adipose tissue function, which all converge on the idea of insulin deficiency rather than the insulin resistance seen in European cohorts. These mechanisms continue to be largely invisible due to current research, which is framed around obesity-centered models from studies done on primarily European populations.

This also may have direct consequences. When diabetic risk tools and treatment guidelines are centered around groups where obesity is the main cause of disease, they systematically sideline risk in populations such as lean South Asians. The genetic variants, metabolic subtypes, and adipose dysfunction patterns described in this review would not have been identified without population-specific study designs. Ultimately, expanding genetic research to include a wide range of communities is not only necessary for understanding T2D, but it is essential to ensure that diagnosis and treatment are accurate across all groups.

Tables and Figures

Table 1. Summary of the four candidate explanations for lean Type 2 diabetes in South Asian populations, the hypothesis tested, key supporting evidence, and the resulting conclusion.

Question	Hypothesis Tested	Key Evidence	Results / Conclusion
Distinct or shared genetics?	Lean T2D is a separate genetic disease, or caused by hidden obesity	Same risk alleles in lean and obese, but stronger effects in lean. Low C-peptide shows insulin deficiency, not resistance. High genetic risk results in earlier onset at lower BMI.	Shared genetics, not distinct. Lean T2D reflects reduced metabolic tolerance to the same variants.
Rare mutations?	Some lean patients carry rare high-impact β-cell mutations missed by GWAS	Exome sequencing finds rare loss-of-function variants with large effects. Monogenic diabetes variants found in 2.4% of youth-onset T2D cases.	Rare β-cell mutations explain some early-onset cases.
Adipose dysfunction?	Polygenic lipodystrophy-related	16 of 138 lipodystrophy genes dysregulated in Indian T2D. South	Lean T2D represents mild polygenic adipose

Question	Hypothesis Tested	Key Evidence	Results / Conclusion
	variants impair fat storage, causing ectopic fat at normal weight	Asians have more ectopic fat even after BMI adjustment. GWAS loci affect fat distribution, not total adiposity.	dysfunction. Not excess fat, but misplaced fat causing metabolic harm.
Thrifty genotype?	Populations evolved efficient energy storage that now causes obesity and insulin resistance	Patients show insulin deficiency, not enhanced storage. SIDD is the dominant subtype (53%). GWAS loci affect β-cell function, not energy storage.	Thrifty genotype predicts obesity + resistance. Data shows leanness + deficiency, the opposite. Hypothesis does not explain the pattern and highlights the need for population-specific studies.

Acknowledgements

Thank you to Dr. Natalie Deforest and Dr. Freddy Nguyen, my mentors, for their help, guidance, and valuable feedback throughout my process of writing this research paper.

Works Cited

Flannick, J., et al. "Exome Sequencing of 20,791 Cases of Type 2 Diabetes and 24,440 Controls." Nature, vol. 570, no. 7759, 2019, pp. 71–76. https://doi.org/10.1038/s41586-019-1231-2.

Hodgson, S., et al. "Genetic Basis of Early Onset and Progression of Type 2 Diabetes in South Asians." Nature Medicine, vol. 31, 2024, pp. 323–331. https://doi.org/10.1038/s41591-024-03317-8.

Kooner, J. S., et al. "Genome-Wide Association Study in Individuals of South Asian Ancestry Identifies Six New Type 2 Diabetes Susceptibility Loci." Nature Genetics, vol. 43, no. 10, 2011, pp. 984–989. https://doi.org/10.1038/ng.921.

Kwak, S. H., et al. "Genetic Architecture and Biology of Youth-Onset Type 2 Diabetes." Nature Metabolism, vol. 6, no. 2, 2024, pp. 226–237. https://doi.org/10.1038/s42255-023-00970-0.

Misra, A., and L. Khurana. "Obesity-Related Non-communicable Diseases: South Asians versus White Caucasians." International Journal of Obesity, vol. 35, no. 2, 2011, pp. 167–187. https://doi.org/10.1038/ijo.2010.135.

Mohan, V. "Lessons Learned from Epidemiology of Type 2 Diabetes in South Asians: Kelly West Award Lecture 2024." Diabetes Care, vol. 48, no. 2, 2025, pp. 153–163. https://doi.org/10.2337/dci24-0046.

Neel, J. V. "Diabetes Mellitus: A ‘Thrifty’ Genotype Rendered Detrimental by ‘Progress’?" American Journal of Human Genetics, vol. 14, no. 4, 1962, pp. 353–362.

Perry, J. R. B., et al. "Stratifying Type 2 Diabetes Cases by BMI Identifies Genetic Risk Variants in LAMA1 and Enrichment for Risk Variants in Lean Compared to Obese Cases." PLoS Genetics, vol. 8, no. 5, 2012, e1002741. https://doi.org/10.1371/journal.pgen.1002741.

Prasad, R. B., et al. "Subgroups of Patients with Young-Onset Type 2 Diabetes in India Reveal Insulin Deficiency as a Major Driver." Diabetologia, vol. 65, no. 1, 2022, pp. 65–78. https://doi.org/10.1007/s00125-021-05543-y.

Salvatore, T., et al. "Current Knowledge on the Pathophysiology of Lean/Normal-Weight Type 2 Diabetes." International Journal of Molecular Sciences, vol. 24, no. 1, 2022, article 658. https://doi.org/10.3390/ijms24010658.

Saxena, A., et al. "The Common Pathophysiologic Threads between Asian Indian Diabetic’s ‘Thin Fat Phenotype’ and Partial Lipodystrophy: The Peripheral Adipose Tissue Transcriptomic Evidence." Adipocyte, vol. 9, no. 1, 2020, pp. 253–263. https://doi.org/10.1080/21623945.2020.1776082.

Siddiqui, M. K., et al. "Young-Onset Diabetes in Asian Indians Is Associated with Lower Measured and Genetically Determined Beta Cell Function." Diabetologia, vol. 65, no. 6, 2022, pp. 973–983. https://doi.org/10.1007/s00125-022-05671-z.

Stefan, N., et al. "Causes, Characteristics, and Consequences of Metabolically Unhealthy Normal Weight in Humans." Cell Metabolism, vol. 26, no. 2, 2017, pp. 292–300. https://doi.org/10.1016/j.cmet.2017.07.008.

Yajnik, C. S., et al. "Polygenic Scores of Diabetes-Related Traits in Subgroups of Type 2 Diabetes in India: A Cohort Study." The Lancet Regional Health – Southeast Asia, vol. 14, 2023, article 100182. https://doi.org/10.1016/j.lansea.2023.100182.

Hormonal Physiology and Performance: How Menstrual-Cycle Tracking Improves Training and Reduces Injury in Female Athletes

By Sai Pragna Prakash[1,2], Sai Pranati Prakash[3]

Abstract

While sports training is increasingly moving toward personalized routines, female athletes still face challenges, as there is limited focus on the menstrual cycle and the impact it has on their performance. There are also many gaps in understanding how female specific factors such as neuromuscular control, body temperature regulation, and ligament stability affect training and performance. This study seeks to understand whether tracking the menstrual cycle in sportswomen and adjusting the intensity and duration of their training accordingly can improve performance, reduce injury risk and lead to better recovery. The research uses a qualitative case study approach, focusing on real-world sports teams and training programs. These include the U.S. Women's National Team, Chelsea FC Women, the Stanford FASTR program, and Project RED-S. It looks at available data on performance and injuries and studies the different tracking methods used by these teams to understand how they manage training schedules and injury risk for their female athletes. Results suggest that tracking the menstrual cycle helps make training more personalized and may reduce some injuries. However, there is no clear, irrefutable evidence of improved performance. There are also practical challenges such as social stigma attached to menstruation, lack of awareness and interest amongst the coaching staff and concerns about data accuracy. Overall, this study infers that menstrual cycle tracking is an important step towards supporting more inclusive and effective training in women's sports, but it works best when it is supported by greater awareness, proper education and regular use.

Key words female athletes, injury risk, menstrual cycle, tracking, sports training and education

Introduction

Female athletes experience non-contact ligament injuries at a rate which is three to six times higher as compared to male athletes.[1,2] One of the factors contributing to this disparity is the role of hormonal changes during the course of the menstrual cycle.[3,4] Fluctuations in estrogen and progesterone during various phases of the menstrual cycle can influence several factors responsible for athletic performance like strength, fatigue, recovery, and neuromuscular control in female athletes.[5,6] Although the participation of women in competitive sports has grown significantly in the past few decades, there is still a lack of high-quality research explaining how the menstrual cycle impacts athletic performance in female athletes.[7] Besides, the hormones-related physical changes during the menstrual cycle make research difficult and complex.[8] Also, most training models are designed around the male physiology, and female

[1] Mercer Island School District, Mercer Island, USA
[2] Mt Baker Rowing and Sailing Center, Seattle, USA
[3] School of Public Health, University of Washington, Seattle, USA

athletes have been underrepresented in sports research with most sports science studies focusing on male athletes.[9] Since hormone patterns vary not only between athletes but also between cycles, individual tracking of the menstrual cycle can become an important tool in developing a most useful training schedule.[10] If these physiological differences are ignored, female athletes may end up training during periods when their bodies are vulnerable and not fully ready, leading to fatigue injury and subpar performance, and increased risk for long term injuries.
Menstrual cycle symptoms can affect athletes both physically and psychologically. Physical symptoms typically include cramps, low energy, heavy bleeding, pain, headaches, fatigue, and muscle soreness. Psychologically, athletes often experience increased anxiety, mood swings, lack of focus, and reduced confidence during certain parts of their menstrual cycle.[11] Although these symptoms frequently disrupt the training consistency and performance of female athletes, athletic programming often overlooks this critical factor in designing athletic training schedules. The menstrual cycle has four main phases: menstrual, follicular, ovulatory, and luteal.[12] (Figure 1). Each phase involves changes in two hormones, estrogen and progesterone. Estrogen is more dominant during the follicular phase and affects muscle function and coordination. Progesterone becomes more dominant during the luteal phase and impacts body temperature, fatigue levels, and cardiovascular responses. These hormone changes help explain why athletes may feel and perform differently at different times of the cycle.

Several studies have observed performance changes across suffering phases of the menstrual cycle.[6] While some trials show that the menstrual cycle significantly alters the performance of female soccer players, other studies do not find a strong correlation.[13] It is often seen that performance dips during the early follicular phase, particularly in high-intensity and endurance sports. In contrast, some studies have shown increases in strength around ovulation due to higher estrogen levels.[14] These mixed findings underscore the need for clearer, more individual athlete focused research. Investigating whether menstrual cycle monitoring can mitigate injury risk and improve athletic performance remains a pivotal area of study.

Methodology

We used a qualitative and comparative research approach to explore the connection between the menstrual cycle and the performance recovery and injury patterns of female athletes. To gain a broader understanding of the topic, the study combined three methods: reviewing existing academic literature, examining menstrual cycle tracking technologies and analysing real life case studies from the sports industry. The purpose of this approach was to get a better understanding on how hormonal changes during the different stages of the menstrual cycle affect women athletes physically and how sports coaches are using this information to develop more personalized and effective training programs for women.

The research began with a structured literature review using peer reviewed journals, research papers and sports science articles. Sources were collected from several databases such as PubMed®, ScienceDirect®, and SpringerLink. Most of the studies selected were published between 2010 and 2025 to ensure that the information was recent and relevant. The review

mainly focused on hormonal fluctuations, neuromuscular control, recovery, injury mechanism and performance changes experienced by female athletes during different menstrual cycle phases. We also examined menstrual cycle tracking technologies used in modern sports training. Wearable devices such as WHOOP and Catapult GPS were studied to understand how they track factors like HRV, sleep, recovery and training load. Platforms such as FitrWoman and Wild.AI were reviewed to examine how menstrual cycle phases are monitored and used for training recommendations. Finally, a comparative analysis of selected case studies was conducted which included examples such as the US Women's National Soccer Team, Project RED-S and the Stanford FASTR program. The case studies were analysed to identify similarities and differences related to athlete readiness, injury prevention, communication between athletes and coaches and the challenges involved in implementing cycle-based training programs in professional sports environments.

Literature Review

Menstrual Physiology & Performance

Among female athletes, the natural menstrual cycle affects several day-to-day training metrics such as heart rate variability (HRV), resting heart rate (RHR), respiratory rate and quality of sleep. The underlying cause for these is the changes in the hormones, estrogen and progesterone. When the level of progesterone rises after ovulation, body temperature increases and the body moves toward a more sympathetic state. This is often described as a "fight-or-flight" state, where resting heart rate, blood pressure, and alertness increase. As progesterone peaks in the late luteal phase, many athletes feel that recovery is harder during this time. This overload caused by the rise of progesterone, can also have long term effects like anxiety, insomnia, and digestive issues.

After menstruation, when progesterone levels drop, the body often returns closer to baseline, and recovery may feel easier compared with the late luteal phase. These patterns are clearly visible in large datasets. Sims et al. found that women with a natural menstrual cycle had higher HR and better recovery at the start of the cycle once menstruation has passed in the early follicular phase.[15] This was followed by a decline in both toward the late luteal phase. RHR and breathing rate also showed a rise across the cycle as progesterone levels increased. Sproll et. al. also described that HRV improves as the cycle moves toward the middle phases.[16] In their analysis, they included the first bleeding days within the "early follicular" phase, which explains why HRV appeared lower at the start of their early-follicular window. After those initial days, their pattern followed the same trend reported by Sims et al. that HRV becomes more stable and tends to improve before declining again later in the cycle. Birth-control hormones follow a different pattern. Athletes using combined oral contraceptives had a different HRV and recovery pattern across the pill cycle.[15] HRV dropped at the start of active pill weeks and then rose again. There was a stronger drop in next-day recovery for the same amount of cardiovascular strain compared to athletes who had a natural cycle. Progestin-only users had patterns closer to natural

cycles. Taken together, these two studies show the effects of hormones, both natural and birth-control related, on how the body responds to training and recovery.[15,16] This helps explain why some phases feel harder than others, and why athletes may not show the same recovery across the entire cycle.

Phase Dependent Variation

Research has shown that performance in female athletes does not remain constant throughout the menstrual cycle. It changes at different points in the cycle. When we examine how elite, highly trained female athletes accelerate and decelerate, there are clear phase-related changes.[17] It has been seen that maximum effort movements usually stay stable, but actions that depend on fine neuromuscular control, such as hard landings, impacts, take-offs, or quick sprints, are likely to vary. These movements require tight coordination between the muscles and the nervous system for the body to stay balanced and absorb force, and this coordination can vary depending on the phase of the cycle. It is difficult to come to a definite conclusion because every athlete shows a different response. Some women athletes reported feeling strongest during the follicular phase, while others performed better around ovulation. These varying experiences again highlight the need for personalised cycle tracking for female athletes rather than trying to apply one universal pattern to everyone. Figure 2 shows the graph depicting ligament laxity peaking pre-ovulation & performance dipping in the late luteal phase.

Injury Mechanisms in Female Athletes

Studies show that rising estrogen levels affect certain ligaments, thereby increasing the risk of anterior cruciate ligament (ACL) injuries. Dos'Santos et al. note that changing estrogen and progesterone levels can reduce stiffness in tendon and ligaments.[18] When these tissues become more flexible, joints may be less stable during movement. These hormonal shifts also cause issues like menstrual dysfunction (irregular or missed cycle) and low energy levels, both of which are associated with a higher injury risk. Neuromuscular control changes across cycle phases affect landing mechanics. It has been seen that injuries went up mainly during the follicular and luteal phases of the menstrual cycle. About 20% of the injuries linked to hormonal changes were "overdue," implying that they occurred after the expected start date of menstruation.[19] The type of injury also differed depending on the phase of the cycle. Martin et. al. found an 88% higher rate of muscle and tendon injuries in the late follicular phase compared with the early follicular phase.[19] Injuries such as muscle tears, strains, cramps, and tendon ruptures were nearly twice as common in the late follicular phase as compared to any other phase. Joint and ligament injuries, on the other hand, were more common during the luteal phase. These injury trends are associated with the rise and fall of ovarian hormones, with changes in estrogen having a key impact on injury risk.

Technology & Tracking Tools

Wearable tools such as WHOOP, Oura, and Catapult are now becoming very popular amongst athletes to track sleep, strain, and global positioning system (GPS) load and other important parameters. Bellenger et al. studied the performance of wearable device data, in particular the WHOOP strap in measuring HRV and heart rate during sleep in level water polo players who were competing at the Olympic level.[20] The study found that the device recorded these metrics fairly accurately and reliably without any major challenges. The study also studied WHOOP's Strain score, which estimates the overall cardiovascular load over the course of the day. It showed that HR and HRV recorded during sleep provide a practical and simple way for athletes and non-athletes alike to monitor the body's recovery levels and readiness for the day. At the same time, it is important to point out that the strain score should be interpreted with a level of caution because it can be influenced by many day-to-day factors such as sleep quality and mental stress levels.

In another study, researchers focused on menstrual-cycle tracking in sportswomen and delved into why it is so important for the performance of female athletes.[21] This study also highlights the fact that there is no standardization of tracking methods and the terminology used in research, which makes it harder to compare results between athletes or across different teams. Even though some progress has been made, there is still no clear universally accepted system which can be confidently used by coaches, athletes and training programs. Table 1 shows the various metrics tracked by wearable devices: HRV accuracy, sleep tracking, cycle prediction and training recommendations.

Gaps in Training Models

There is still no clear standard method of adjusting training around the menstrual cycle, and most sports teams including the elite teams use the cycle tracking information to interpret in their own different ways, which also was highlighted in the recent UEFA-led consensus statement: the science is uneven, and teams are often unsure about what to measure or how to actually use the data to improve the training.[22] The consensus statement nevertheless, makes an attempt to give evidence-based guidance on what should be tracked and how those measurements can support an athlete throughout the cycle. Even with these recommendations, the authors note that many top programs still do not formally include menstrual-cycle information in their load-management or planning models.[22] The statement also emphasizes that cycle tracking should be used not only to improve performance but also to support the athlete's overall health. A consistent, standardized approach would make it much easier for coaches, trainers, and clinicians to understand how hormonal changes affect the way an athlete responds to different types of training.

Results

Case Study 1 — Chelsea FC Women

Background

Chelsea FC Women were one of the first football clubs to give serious consideration to and invest in cycle and symptom tracking and to use the collected information as a part of everyday training of female players. They entered into a partnership with the menstrual cycle tracking app called FitrWoman to create customized individual plans for each player based on the phases of their menstrual cycle. The ultimate objective was to lower soft-tissue injury risk by understanding how the different phases of their menstrual cycle may affect how they feel, how they recover and how ready they are to train. As a result, they could take steps to prevent injuries rather than react to them once they had already occurred.

Data Collected

The female athletes recorded their symptoms such as cramps, fatigue, mood changes, bloating, sleep quality headaches etc and the phase of their menstrual cycle through the FitrWoman app. This information helped staff understand how different phases of the cycle impacted an athlete's training and recovery. The data collected was combined with GPS training data such as distance covered, speed, accelerations and strength indicators like jump height and sprint performance. Examining menstrual symptoms alongside these physical metrics helped staff spot when an athlete might be at a higher risk for soft-tissue injuries like ACL strain, which can be influenced by menstruation linked hormonal changes.

Key Findings

Chelsea FC Women found that there was a reduction in soft-tissue injuries after they started planning training schedules around player cycles. Chelsea reported that players often showed lower readiness in the early follicular phase and that ovulation was linked with signs of increased ligament laxity which can raise ACL risk around ovulation. This allowed them to adjust training demands to make them more safe.[23,24] Players said that they felt more supported as they had a better understanding of what was happening in their bodies and their training schedules could be tweaked when needed.

Implications

Women players explained that learning about their cycles helped them make better choices about recovery and training and helped them perform the best throughout the month regardless of the phase they are in. This case reinforces that adjusting training intensity to each player’s menstrual-cycle phase is not only helpful but necessary. It also highlights why personalised custom training approaches work better than giving all athletes the same plan.

Disclaimer

Chelsea FC did not share any exact percentages or numerical data related to injury-reduction or performance. All publicly available sources give general qualitative trends without any specific figures.

Case Study 2 — WHOOP Hormonal Analytics

Background

WHOOP collects everyday health related data like HRV, RHR, sleep, and daily strain from a multitude of people who wear the device.[25] While conducting its menstrual-cycle research, the company studied more than 45,000 menstrual cycles from about 11,000 members, giving better understanding of how the body changes across different phases. The group of users included in this study were not restricted to elite athletes. It included a mix of differing lifestyles and training backgrounds including professional and recreational athletes, fitness enthusiasts and everyday users who simply wear the device to track their health parameters. Instead of merely tracking menstruation dates on a calendar, WHOOP does a deep dive into how the body responds to hormonal changes across the cycle in the course of the cycle.[26] This approach led to the emergence of a new measure called cardiovascular amplitude, which follows the rise and fall of HRV and RHR throughout the cycle.

Data Collected

WHOOP combines data received from the wearable sensor and self-reported symptoms by the users themselves to study the impact of the menstrual cycle on the female body. The device itself automatically keeps track of important parameters like nightly HRV, RHR, breathing rate, and sleep quality. The data received goes into creating the daily recovery score which shows how prepared a person's body is to undertake physical or mental effort the next day. To supplement this score, users also add their own notes like where they are in their menstrual cycle and symptoms they might feel such as fatigue, mood changes, cramps, headaches, bloating, sleep difficulties, or heaviness in the legs. When WHOOP combines device data with symptom logs entered by the user, it becomes easier to understand how hormonal changes appear in daily life and how they impact energy levels, recovery, and overall performance in women. (WHOOP, 2025a).

Key Findings

WHOOP noticed some clear patterns across the various phases of the menstrual cycle.[27,28] The early follicular phase, which is the beginning of menstruation, most often shows the lowest readiness scores. HRV tends to dip during this time, and RHR also shifts usually towards the lower side as the body tries to adjust. Around ovulation, WHOOP found small, short-term fluctuations in HRV that point to temporary dips in the body's stability. During the luteal phase, heart related patterns were steadier and typically showed higher RHR and lower HRV compared to the earlier phases. People also reported feeling fatigued and recovering more slowly during this phase of the cycle. These findings suggest that training and workouts routines cannot be the same every day. Their intensity needs to be adjusted based on readiness scores to a level that the body can handle without undue stress.

Implications

WHOOP's findings match with what existing research has shown: hormonal changes can affect recovery, stability and fatigue. In this model, data comes from real people in normal everyday routines instead of a controlled environment. Instead of looking at low-readiness days as a sign that they lack discipline or are not working hard enough, athletes can connect these changes to the body's normal responses to hormonal fluctuations. This awareness helps women athletes plan better, avoid stressing about their bodies on low readiness days and reduce the overall risk of overexerting and injury. Overall, WHOOP's work shows how wearables can help women train in a manner which is in tune with their bodies.

Disclaimer

WHOOP reports overall trends from its data but does not provide exact injury-risk percentages or detailed statistical reports. All results come from aggregated wearable data from their website.

Case Study 3 — Global ACL Injury Patterns

Background

One topic which is often discussed among injury risks specific to women athletes is the high rate of ACL injuries seen in women's sports, with many studies showing a 2 to 6 times higher risk of non-contact ACL injury than male athletes especially in sports like soccer and basketball.[18] The connection between the menstrual cycle and ACL injury is not completely clear as there are great differences amongst female athletes themselves in terms of their training schedules, hormonal patterns and contraceptives use.

Data Collected

Large epidemiological datasets from National Collegiate Athletic Association (NCAA) and International Association Football Federation (FIFA), provide valuable information about the timing and frequency of ACL and other knee injuries in competitive sport.[29,30] Most injury databases typically record injury type, timing, activity being performed, and athlete exposure. When biomechanical analysis is combined with hormonal studies, researchers can understand how the cycle related hormonal fluctuations in women influence tissue laxity and movement during sports training and activity.

Key Findings

Across these biomechanical and epidemiological studies, there is a clear pattern that links the menstrual cycle phases with ACL injury and other risk indicators.[19] Multiple systematic reviews have shown that the pre-ovulatory or ovulatory phase, which is marked by rising estrogen levels, often correlates to peak ACL injury risk in female athletes. This is generally days 9-14 of a 28 day cycle and is when there are most data on the number of non-contact ACL

incidents occurring. Absolute causal data is difficult to find, however there is enough proof to point to the association of hormonal changes and neuromuscular control during this phase.

Implications

These findings offer valuable insights for female athletes and their coaches to plan their training schedules in a manner which will help them avoid injuries. If it is known that there is a higher risk of ACL injury in the ovulatory phase, then coaches and medical staff could adjust the training intensity in that period to make it safer. They can also keep a close watch on the athlete and reduce specific high-risk movements like repeated jumping or rapid pivoting during these specific phases. It will also avoid situations where athletes are pushed into intense training when the body is more vulnerable and not ready for it. Neuromuscular training programs like FIFA 11+ have shown an impressive reduction of 30 to 50% in ACL injury rates in female athletes.[31] While large-scale research studies are still needed, it has become clear that adjusting training loads and exercise schedules around an athlete's menstrual cycle has the potential to lower ACL injury risk and promote long term health and safety for sportswomen.

Case Study 4 — NCAA Women's Teams tracking athletes' cycles

Background

Many athletic programs are now paying more attention to how a female athlete's cycle relates to and influences their training and performance. Recently, teams in the NCAA such as soccer, track, basketball, and volleyball have all begun tracking their athletes' cycles to learn how hormones such as estrogen and progesterone are affecting their injury risk and fatigue as they move through the season.[29] To give context, in the United States alone, more than 220,000 women compete in NCAA sports each year, which makes this case study notable in the context of monitoring how female athletes train and recover during various phases of their menstrual cycle.[32] Coaches can use this information to detect and follow patterns, and help their athletes perform to the best of their ability.

Data Collected

One of the most important sources of data comes from the NCAA's GPS load and heart rate zones recorded during practices and games and tracking female athlete's cycles. This data gives insight into how much the body is in stress. For example, in soccer research, athletes' movements during matches were recorded using GPS to see how their hearts responded to the distance they were covering across different days. In addition to this data, athletes also personally record factors like their symptoms, soreness, sleep quality, and their cycle phase. They can report symptoms like tiredness, cramps, headaches, and difficulty in sleeping on different days of their cycles. They also mark their menstruation days and other phases of the cycle. All this data is given to researchers who then compare the reported information with the tracked information to identify patterns. One study of a volleyball team found that there were

nearly 12 symptoms that were reported per cycle, which included difficulty sleeping and tiredness.[33] Scientists also track game performance metrics to see if there are links between cycle phase and how athletes perform in competition.

Key Findings

Research on menstrual cycle tracking in athletes has revealed that there are consistent patterns. More injuries, especially joint and ligament, occurred during the luteal phase of the cycle. Higher levels of tiredness and worse sleep quality was also reported during the early and late luteal phases. This is also when athletes reported more tiredness and worse sleep. These patterns give teams information about when athletes may need more recovery and rest.

Implications

The research suggests that tracking menstrual cycles is an important tool for athletes and coaches. Without needing any fancy equipment, coaches and athletes can pay attention to cycle phases and notice patterns in how they feel from week to week. Athletes record days when they feel certain symptoms and use this information to decide when their body needs more recovery versus when to train harder. Even without advanced technology and monitoring systems, consistent and accurate symptom tracking can improve training decisions and prevent overuse, injury, and burnout. Table 2 compares teams and programs that use cycle tracking, like Chelsea women, NCAA teams, and WHOOP athletes.

Discussion

Patterns Across All Case Studies

All the studies show that readiness and injury risk among female athletes are impacted by and linked to the various phases of the menstrual cycle. Athletes note decreased readiness and HRV specifically around certain phases, and although variable, the fluctuation of these sex hormones can greatly impact recovery and overall health.[5,15] Several studies suggest that ovulation may be a period of increased injury risk. This is due to the increase of estrogen in the body, which increases ligament laxity, and overall risk of soft-tissue injury.[4,18] However, it is important to clearly mention that not all studies report a similar pattern, and there are significant differences between individual athletes. The use of wearable monitoring devices that track markers such as HRV, sleep, and recovery has made it easier to observe how these affect athletes during their everyday training schedules.[15] As these tools constantly track data from real world activities, they provide information and insights which might be missed in controlled laboratory settings.[11]

Effectiveness of Cycle-Based Training

Cycle-based training has proven to be very effective in helping female athletes better understand their bodies and accordingly adjust their training schedules. Training programs that

have been tailored to specific phases of an individual athlete's cycle have been linked to better injury prevention while keeping athletes consistent in training.[4] This reflects the transition from fixed training schedule to more adaptable, athlete-cantered programs. By tracking their menstrual cycles, athletes can better understand how hormonal changes impact their strength, energy levels, and sleep, and how these factors affect their training. Figure 3 gives a schematic overview of how menstrual-cycle tracking tools convert physiological data into training insights.

Barriers and Limitations

One of the major challenges that comes in the way of implementing cycle-based training programs is the limited education amongst athletes regarding female physiology and the basic processes of the menstrual cycle. For example some of the women athletes from the Chelsea FC Women claim that they had very limited awareness of the monthly hormonal changes occurring in their bodies and its profound impact on their overall health and performance.[24] The social stigma around the menstrual cycle along with limited education continues to be a major barrier in the way of implementing cycle-aware training schedules. Cycle-based training still needs to be individualized and can be done so using wearable tracking technology like WHOOP sensors, but the data quality is highly variable. Factors like irregular tracking, reliance on self-reported data, and differences in device accuracy can affect the reliability of the results. Apps like FitrWoman, try to address this gap by combining wearable data with self-reported symptoms from athletes to improve the accuracy of readiness assessments.[23]

Implications for Coaching and Sports Science

Research increasingly supports the use of gender-specific training models that take into consideration the physiological differences between male and female athletes. The traditional training schedules are largely developed keeping male physiology in mind as most of the participants are male athletes. As a result, the hormonal changes experienced by female athletes due to their menstrual cycle are not adequately accounted for. Conversely, when training programs are guided by menstrual cycle awareness, it leads to mutual benefits both for the coaches and for the women undergoing the training. By combining performance tracking tools with self-reported symptoms logged into an app, coaches can produce a tailored workout and training schedule for every female athlete that supports her training capabilities. The data received from wearable trackers for important metrics like HRV, sleep and recovery can help recognize patterns in the body’s readiness and fatigue levels. This information helps coaches adjust training intensity, duration and recovery time in a manner which supports the athlete in improving performance and reduces the risk of potential injury and overtraining. Figure 4 gives a conceptual roadmap illustrating future developments in menstrual-aware athlete monitoring.

Conclusion

Menstrual cycle tracking can improve both the safety and performance of female athletes. More than 75% of female athletes experience negative side effects of their menstrual cycle on

their athletic training including pain, fatigue and greater risk for soft tissue injury. Also, hormonal changes have an impact on performance as they affect factors like sleep, mood, recovery and energy levels. Many sports teams have started using wearable devices to better manage these symptoms. Wearables like WHOOP and FitrWoman track heart rate, recovery and cycle related changes making it easier to identify patterns and customize training in the interest of overall health and safety. Tracking the menstrual cycle in female athletes has proven to be useful. However, utilization of cycle-based tracking is still limited. Responses received from the tracking of the menstrual cycle have high interpersonal variability: one-size-doesn't-fit-all. Nevertheless, with increasing female participation in elite level sports, the availability of data related to physiological phases has improved. For example, WHOOP has been able to track 45,000 cycles from 11,000 users, helping researchers identify trends in how hormonal changes affect the body and performance. Further research in this area can help create customized individual training plans, influence how athletes feel physically and mentally, break down stigma as well as encourage better understanding and education. Most importantly, it can inspire a new generation of female athletes to better understand their bodies and to make informed decisions about their training and recovery.

Works Cited

1.Ireland ML. The female ACL: why is it more prone to injury? *Orthop Clin North Am*. 2002;33(4):637-651. doi:10.1016/s0030-5898(02)00028-7

2.Wojtys EM, Huston LJ, Lindenfeld TN, Hewett TE, Greenfield ML. Association between the menstrual cycle and anterior cruciate ligament injuries in female athletes. *Am J Sports Med*. 1998;26(5):614-619. doi:10.1177/03635465980260050301

3.Myers SJ, Knight RL, Wardle SL, et al. Effect of Menstrual Cycle and Hormonal Contraception on Musculoskeletal Health and Performance: Protocol for a Prospective Cohort Design and Cross-Sectional Comparison. *JMIR Res Protoc*. 2024;13:e50542. doi:10.2196/50542

4.Fort-Vanmeerhaeghe A, Pujol-Marzo M, Milà R, et al. Injury Risk and Overall Well-Being During the Menstrual Cycle in Elite Adolescent Team Sports Athletes. *Healthcare*. 2025;13(10):1154. doi:10.3390/healthcare13101154

5.Jones BP, L'Heveder A, Bishop C, et al. Menstrual cycles and the impact upon performance in elite British track and field athletes: a longitudinal study. *Front Sports Act Living*. 2024;6:1296189. doi:10.3389/fspor.2024.1296189

6.Cabre HE, Joniak KE, Ladan AN, et al. Effects of Hormonal Contraception and the Menstrual Cycle on Maximal Strength and Power Performance. *Med Sci Sports Exerc*. 2024;56(12):2385-2393. doi:10.1249/MSS.0000000000003524

7.McNulty KL, Elliott-Sale KJ, Dolan E, et al. The Effects of Menstrual Cycle Phase on Exercise Performance in Eumenorrheic Women: A Systematic Review and Meta-Analysis. *Sports Med*. 2020;50(10):1813-1827. doi:10.1007/s40279-020-01319-3

8.Myths and Methodologies: Reducing scientific design ambiguity in studies comparing sexes and/or menstrual cycle phases. doi:10.1113/EP086797

9.Ose BM, Eisenhauer J, Roepe I, Herda AA, Vopat BG, Vopat LM. Where Are All the Female Participants in Sports and Exercise Medicine Research? A Decade Later. *Am J Sports Med*. 2025;53(8):2022-2028. doi:10.1177/03635465241278350

10.Solli GS, Sandbakk SB, Noordhof DA, Ihalainen JK, Sandbakk Ø. Changes in Self-Reported Physical Fitness, Performance, and Side Effects Across the Phases of the Menstrual Cycle Among Competitive Endurance Athletes. *Int J Sports Physiol Perform*. 2020;15(9):1324-1333. doi:10.1123/ijspp.2019-0616

11.Chica-Latorre S, Knight-Agarwal C, McKune A, Minehan M. A systematic review and meta-aggregation of the experiences and perceptions of menstrual and hormonal contraceptive cycle-related symptoms in female athletes. *J Sci Med Sport*. 2025;28(12):969-979. doi:10.1016/j.jsams.2025.06.005

12.Messinis IE, Messini CI, Dafopoulos K. Novel aspects of the endocrinology of the menstrual cycle. *Reprod Biomed Online*. 2014;28(6):714-722. doi:10.1016/j.rbmo.2014.02.003

13.Machado M, Gonçalves B, Sousa A, Abade E, Nakamura FY, Coutinho D. How does the menstrual cycle affect female football players' performance during small-sided games? *Int J Sports Sci Coach*. 2026;21(1):131-142. doi:10.1177/17479541251366436

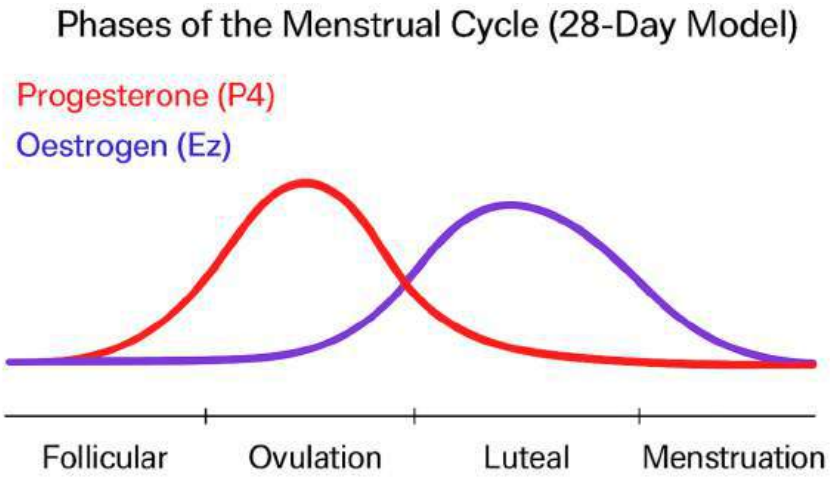
14.Shahid W, Noor R, Bashir MS. Effects of exercise on sex steroid hormones (estrogen, progesterone, testosterone) in eumenorrheic females: A systematic to review and meta-analysis. *BMC Womens Health*. 2024;24:354. doi:10.1186/s12905-024-03203-y
15.Sims ST, Ware L, Capodilupo ER. Patterns of endogenous and exogenous ovarian hormone modulation on recovery metrics across the menstrual cycle. *BMJ Open Sport Exerc Med*. 2021;7(3):e001047. doi:10.1136/bmjsem-2021-001047
16.Sproll M, Otterbach N, Zech A. Neuromuscular Control and Motor Performance Across the Menstrual Cycle in Physically Active Young Females. *Eur J Sport Sci*. 2026;26(5):e70174. doi:10.1002/ejsc.70174
17.Sproll M, Otterbach N, Zech A. Neuromuscular Control and Motor Performance Across the Menstrual Cycle in Physically Active Young Females. *Eur J Sport Sci*. 2026;26(5):e70174. doi:10.1002/ejsc.70174
18.Dos'Santos T, Stebbings GK, Morse C, Shashidharan M, Daniels KAJ, Sanderson A. Effects of the menstrual cycle phase on anterior cruciate ligament neuromuscular and biomechanical injury risk surrogates in eumenorrheic and naturally menstruating women: A systematic review. *PloS One*. 2023;18(1):e0280800. doi:10.1371/journal.pone.0280800
19.Martin D, Timmins K, Cowie C, et al. Injury Incidence Across the Menstrual Cycle in International Footballers. *Front Sports Act Living*. 2021;3:616999. doi:10.3389/fspor.2021.616999
20.Bellenger CR, Miller D, Halson SL, Roach GD, Maclennan M, Sargent C. Evaluating the Typical Day-to-Day Variability of WHOOP-Derived Heart Rate Variability in Olympic Water Polo Athletes. *Sensors*. 2022;22(18):6723. doi:10.3390/s22186723
21.Dole A, Beaven M, Sims ST. Menstrual Cycle Tracking in Sports Research: Challenges, Progress, and Future Directions. *Physiologia*. 2023;3(4):598-610. doi:10.3390/physiologia3040044
22.Verhagen E, Ferrer E, Antero J da S, et al. UEFA consensus statement on menstrual cycle tracking in women's football. *BMJ Open Sport Exerc Med*. 2025;11(3). doi:10.1136/bmjsem-2025-002769
23.A pioneering relationship: How Orreco and Chelsea FC Women work together. March 18, 2022. Accessed July 21, 2026. https://www.chelseafc.com/en/news/article/a-pioneering-relationship--how-orreco-and-chelsea-fc-women-work-
24.Chelsea Women tailor training to players' menstrual cycles. February 14, 2020. Accessed July 21, 2026. https://www.chelseafc.com/en/news/article/chelsea-women-tailor-training-to-players-menstrual-cycles
25.How WHOOP Works | Health Monitoring, Sleep Tracking, Recovery Insights. WHOOP. Accessed July 21, 2026. https://www.whoop.com/us/en/how-it-works/

26.Empowering Your Health: New WHOOP Research Reveals Key Insights on the Menstrual Cycle. WHOOP. March 7, 2025. Accessed July 21, 2026. https://www.whoop.com/us/en/thelocker/new-whoop-research-on-the-menstrual-cycle/
27.Benefits of Working Out on Your Period. WHOOP. May 8, 2026. Accessed July 21, 2026. https://www.whoop.com/us/en/thelocker/why-you-should-work-out-during-your-period/
28.Sleep & Training Based on Menstrual Cycle | WHOOP Podcast. WHOOP. May 21, 2026. Accessed July 21, 2026. https://www.whoop.com/us/en/thelocker/podcast-150-training-sleeping-menstrual-cycle/
29.NCAA Injury Surveillance Program. NCAA.org. Accessed July 21, 2026. https://www.ncaa.org/what-we-do/health-safety-and-performance/injury-surveillance-program/
30.Inside FIFA. Accessed July 21, 2026. https://inside.fifa.com
31.Bizzini M, Dvorak J. FIFA 11+: an effective programme to prevent football injuries in various player groups worldwide-a narrative review. *Br J Sports Med*. 2015;49(9):577-579. doi:10.1136/bjsports-2015-094765
32.NCAA Statistics. Accessed July 21, 2026. https://stats.ncaa.org/
33.Roffler A, Fleddermann MT, de Haan H, Krüger K, Zentgraf K. Menstrual cycle tracking in professional volleyball athletes. *Front Sports Act Living*. 2024;6:1408711. doi:10.3389/fspor.2024.1408711

Figures and tables

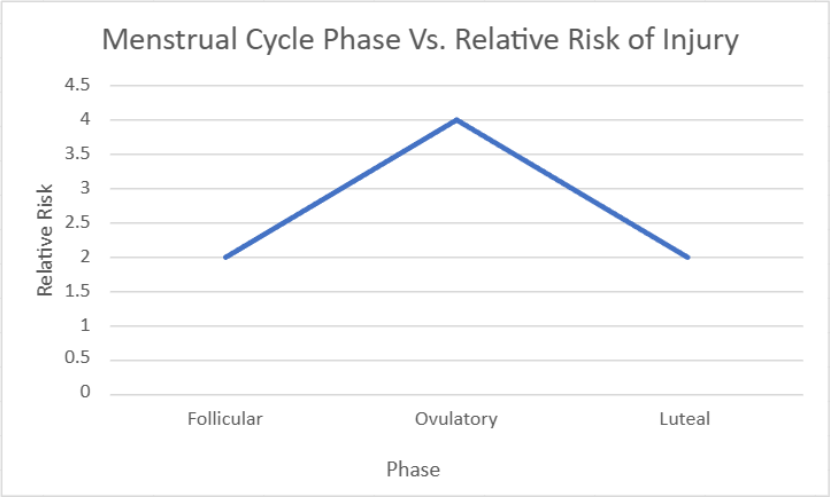
Figures:

Figure 1. Phases of the Menstrual Cycle



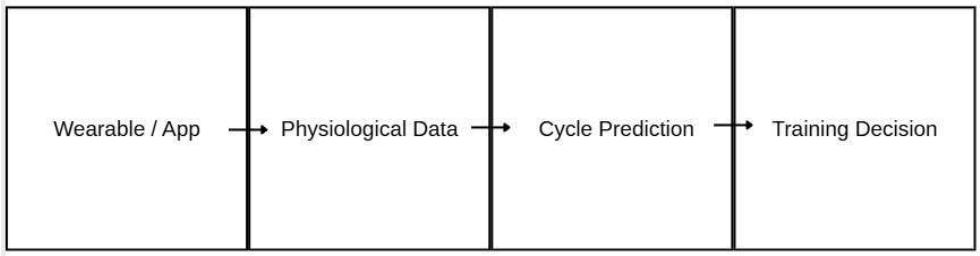
This diagram depicts the hormone level fluctuations across the four distinct phases of a 28-day menstrual cycle model. The red line peaks distinctly during the ovulation phase, while the purple line reaches its highest concentration during the luteal phase.

Figure 2. ACL Injury Risk Timeline Across the Cycle



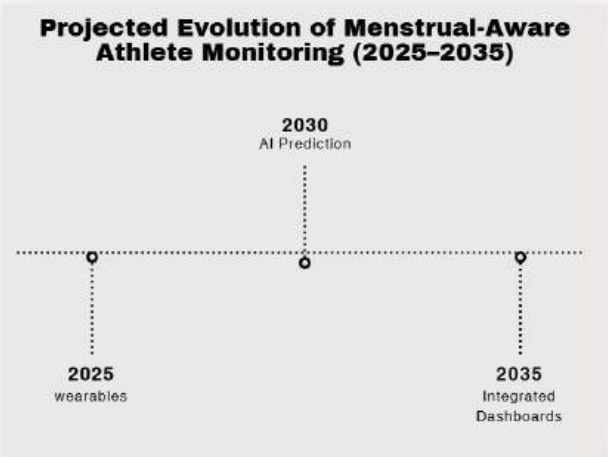
This line graph highlights the changes in the relative risk of athletic injury during different phases of the menstrual cycle. The data indicates that the relative risk remains at a consistent level of 2.0 during both the follicular and luteal phases. Notably, the risk peaks significantly during the ovulatory phase, doubling to a maximum relative risk value of 4.0.

Figure 3. Workflow of Menstrual-Cycle Tracking Systems in Sport.



This flowchart demonstrates the sequential data pipeline for menstrual-aware athletic tracking, moving from hardware/software inputs to final actionable insights. It shows how input from a wearable or app is used to gather physiological data, which then can inform cycle predictions to ultimately guide an athlete's training decisions.

Figure 4. Projected Evolution of Menstrual-Aware Athlete Monitoring (2025–2035).



This timeline shows the projected ten-year evolution of menstrual-aware athlete monitoring technology from 2025 to 2035. It highlights a technological progression from basic wearables in 2025, to AI-driven predictions by 2030, with eventual fully integrated dashboards by 2035

Tables:

Table 1. Comparison of Tracking Tools (WHOOP, Oura, FitrWoman, Wild.AI No)

Technology	HRV* Tracking	Sleep Tracking	Cycle Prediction	Training Guidance
WHOOP	Yes	Yes	Yes	Yes
Oura Ring	Yes	Yes	Partial	Limited
FitrWoman	No	No	Yes	Yes
Wild.AI No	Partial	Yes	Yes	NA

Table 2. Comparison of Sports Programs Using Menstrual-Cycle Tracking in Sport

Team/Program	Level	Technology Used	Training Adjustments	Injury Outcomes
Chelsea FC Women	Elite	FitrWoman	Load modification	Reduced soft-tissue injuries
WHOOP Athletes	Multi-Sport	WHOOP	Readiness-based	Reduced overtraining
NCAA Teams	Collegiate	Surveys/Wearables	Recovery focus	Reduced overuse injuries

FBN-1 Driven Phenotypic Changes in Cardiac Tissue and the Aortic Wall in Patients with Marfan Syndrome By Tanvi Nulkar

Abstract

Marfan syndrome is a genetic disorder that affects individuals globally with a reported prevalence of 1 in 5000 people and incidence of 1 in 3000-5000 (Takeda et al.). The syndrome is classified by the presence of significant cardiovascular and ocular issues. To date, significant advancements in research on Marfan syndrome diagnostic criteria and development of therapeutic interventions have resulted in nearly doubling the expected lifespan of those diagnosed with Marfan syndrome shifting the range from 32-48 years to 70-80 years of life expectancy.(Salik and Rawla) This change in life expectancy comes from the identification of the primary gene underlying Marfan syndrome pathology, , FBN-1, encoding for fibrinillin-1. This discovery provided an opportunity for formulation of targeted pharmacological therapeutics designed to inhibit the effects of the mutated gene. Mutations in the FBN-1 gene contribute to phenotypic changes in the heart, specifically the cardiac tissue and the aortic wall. These alterations induce significant changes in the cellular infrastructure causing the inhibition of proper distention and recoil during the cardiac cycle. In this paper, I review FBN-1 mutations and their effect on the cardiovascular system and further explore how the impacted vascular system affects the body systemically. .

Introduction

Marfan syndrome was first discovered in 1986 by French pediatrician Dr. Antoine Marfan, after encountering a female patient with abnormally long axial limbs. In 1986 the disease was officially defined and given official diagnostic criteria deemed the Ghent nosology which includes symptoms such as ectopia lentis, a condition causing displacement of the eye's lens in addition to increased axial length (Ghoraba et al.). The Ghent nosology was revised to include emergent discovery of underlying associated gene involved FBN-1, and this revised nosology remains the diagnostic criteria as of 2025. However, while Marfan syndrome causes characteristic external changes to patients, the most problematic symptoms of the disease occur internally, with patients suffering from a wide range of ocular, skeletal, and cardiovascular symptoms (Loeys et al.) as listed in Table 1.

Table 1. List of symptoms and physical effects of patients with Marfan syndrome broken up into organ systems. (Information sourced from: (Loeys et al.))

Organ	**Effect**
Cardiovascular	Dilation of the aorta, dissection, aneurysms, prolapse, regurgitation
Ocular	Ectopia lenis, high myopia, dislocated lenses, retinal detachment

Skeletal	Scoliosis or kyphosis, inward or outward growth of breastbone, slender limbs, flat feet, claw or hammer toes
Other	Spontaneous pneumothorax collapse, dural ectasia, tarlov cysts, nerve pain

Marfan syndrome is a connective tissue disorder that is predominantly hereditary with 75% of cases associated with genetic aberrancy. Hereditary Marfan syndrome is autosomal dominant, meaning that the mutation can be inherited from a single parent. The disorder can also be *de novo* or spontaneous accounting for 25% of cases, particularly in those without previously known hereditary connective tissue disorders. Annually, there are an estimated 1/5000-1/10000 number of new cases worldwide (Takeda and Komuro).

Mutations in fibrillin-1 (FBN-1) account for 93% of Marfan syndrome cases in humans with the two primary variations being haploinsufficient or dominant-negative. The other 7% of cases are caused by mutations in the transforming growth factor beta 1 (TGFβ1) and 2 (TGFβ2), two proteins that are critical for tissue repair (Dietz).

Here I will perform a comprehensive investigation on research findings to date on Marfan syndrome. Specifically this research will focus on those mutations which are specifically derived from haploinsufficient and dominant-negative mutations in the FBN-1 gene to elucidate the differential impacts of how heterogeneic mechanisms result in divergent pathology and pathogenesis of Marfan's disorder.

Fibrillin, encoded by fibrillin-1 is a calcium binding protein found in the extracellular matrix and the basement membranes, where it plays a key role key role in maintenance of connective tissue elasticity, intracellular support, regulation of cell growth, cellular processing, and calcium homeostasis. For example, this can be seen through the link between fibrillin and cardiomyocytes (CM). CMs rely heavily on fibrillin to create structure that they can bind to on the ECM. Fibrillin acts as a bridge that helps to relay signals produced by the cardiomyocytes, often helping to reduce the time it takes for the signals to be relayed (Aalders et al.). In patients that lack adequate production of fibrillin, studies show a noticeable stiffness in CMs. This contributes to the characteristic stiffness of Marfan syndrome. Furthermore, in cultures where there was an instance of mutated fibrillin, CMs were shown to be less abundant when exposed to dynamic strain, the same stress that would be found in the aortic walls. This thereby hints towards reduced strength and resilience of Marfan syndrome CM when exposed to normal stress in cardiovascular tissue. Thus demonstrating that FBN-1 and its protein, fibrillin, play a critical role in the strength of cardiovascular tissue.

Specifically, in the case of aortic tissue, strength is characterized by elasticity, stiffness, compliance (how much the diameter of the aorta changes in response to pressure), viscoelasticity (the rate and load of deformation and recoil), and Anisotropy (the ability of the tissue to share mechanical load across the wall) (Knorst). However, patients with Marfan syndrome often display low compliance, largely due to aortic root and pulmonary trunk dilations, low viscoelasticity due to stiffness of the aortic wall, and low anisotropy due to disrupted architecture in the cardiovascular system (Wozniak-Mielczarek et al.). Thus, in patients with Marfan

syndrome, mutations in the FBN-1 gene directly result in severe cardiovascular challenges. This is because the mutated FBN-1 gene and disordered fibrillin production have a direct impact on the supporting architecture of the organ and on the ability for cells to properly dissipate the force of distention.

The exact mechanism of how the FBN-1 mutation affects the cardiovascular cellular architecture is dependent on the type of mutation variation. The two most common types of FBN-1 mutations can be classified as dominant-negative or haploinsufficient (Groeneveld et al.). Dominant-negative mutations are defined as the malfolding or malfunctioning of FBN-1 and fibrillin. In contrast, haploinsufficient mutations are categorized by the deletion of FBN-1, degradation of mutated fibrillin, or nonsense mediated decay by degradation of FBN-1 mRNA (Takeda et al.).

However, regardless of how the gene is perturbed, all mutations in FBN-1 disrupt the transforming growth factor beta signalling (TGF-β) pathway, causing it to go into overdrive. This results in increased production of growth proteins, contractile proteins, dysregulation of cell growth and proliferation, as well as hyperactivation of the mammalian target of rapamycin (mTOR) pathway through integrins. This downstream dysregulation of proteins and cellular components reduces the integral strength of the cardiovascular tissue and muscle, leading to individuals with Marfan syndrome to be significantly more prone to cardiovascular failures including dissections, aneurysms, and prolapses (Zeigler et al.). However, one detail to note is that in most Marfan syndrome cases an additional pathway, the rapamycin pathway, is also hyperactivated. The primary role of rapamycin is as an mTOR inhibitor where it helps to regulate the pathways promoted by TGF-B (Miyakawa et al.). This highlights a point of investigation that will be discussed in this paper.

Marfan syndrome is widely accepted by the medical community as arising from a combination of overexpression of TGF-β, disrupted collagen architecture, and interference of microfibril assembly (Crosas-Molist et al.). However, the mechanisms connecting FBN-1 mutations to the phenotypic changes in cardiovascular tissue and muscle have not been fully elucidated. Through this research I will be investigating the link between the FBN-1 and the TGF-β pathway, alteration of mTOR signalling, disrupted protein production, to review how these components contribute to cardiovascular failure in patients with Marfan's syndrome.

Although recent advancements in medicine have extended the life expectancy of individuals with Marfan syndrome from 32-48 in the 1970s to 70-80 as of 2025, many of the methods used to treat these individuals are highly invasive and often decrease 'quality of life', or their overall well being (Silverman et al.). In this paper, I aim to highlight the mechanisms in which FBN-1 mutations affect patients, and through that potential areas of investigation for treatments that would be less invasive.

Methodology

Here, I used a series of online databases for data acquisition and performed a comprehensive literature review.. The databases I used include the following in Fig 1:

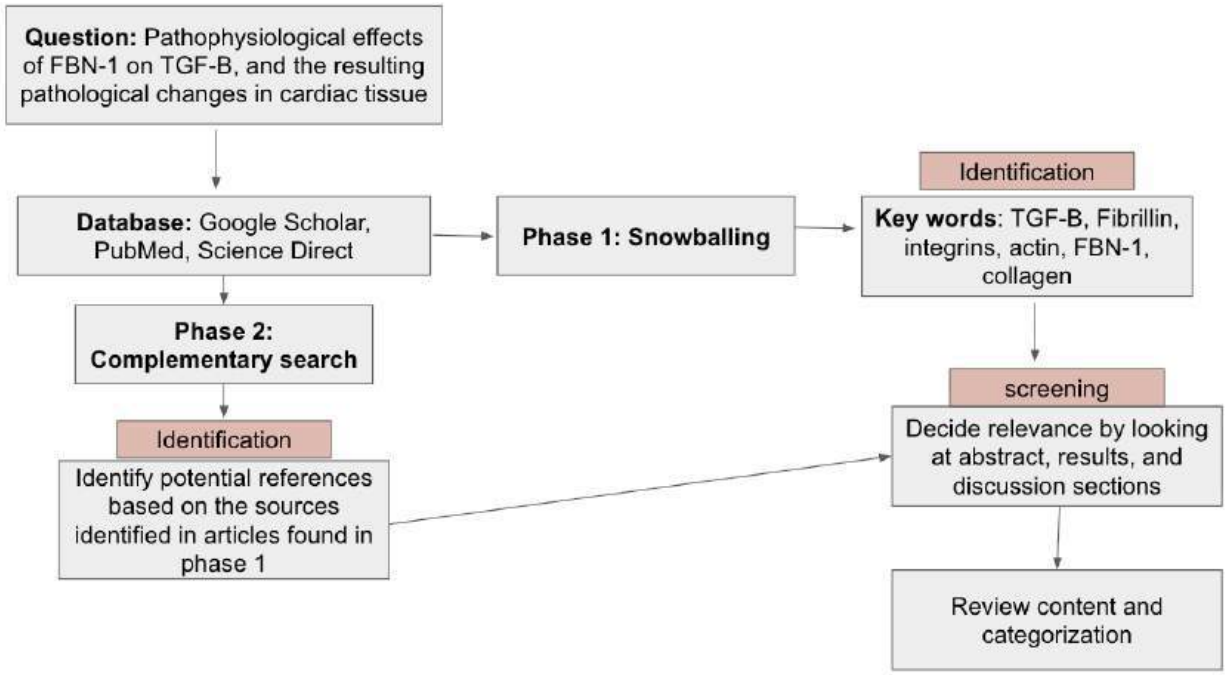


Fig 1. Research Study Design. Demonstrates the snowballing and research process used to identify credible and relevant sources to guide the conclusions made in this paper. Key phrases used to identify sources during phase 1 of the snowballing process were: TGF-B, fibrillin, integrin, actin, FBN-1, and collagen. Furthermore, phase 2 (complementary search), was conducted by referencing the articles cited in relevant papers that were identified in phase 1.

Using the databases shown in Fig. 1, I searched for papers that described results regarding both the molecular composition of the contractile proteins such as collagen, fibronectin, the proteoglycans, integrin, perlecan as well as papers that looked for the link between TGF-β and the protein production. Due to the topic of interest, there are few papers that directly connect FBN-1 with the affected aforementioned proteins. Thus, the aim of this research is to search for a causational relationship between the signaling pathway and regulation impact on proteins using the Snowball method. The Snowball method refers to finding relevant literature by using the references on the first few resources found. Keywords used were: snowball method, TGF-β, FBN-1, collagen, integrin, actin fibers.

Results

Phenotypic changes in vascular smooth muscle cells in patients with FBN-1 Marfan syndrome

FBN-1 driven dysregulation of TGF-B and subsequent mTOR pathway dysregulation causes abnormal collagen architecture, perclan function, fibronectin structure, and microfibril organization. The abnormal and excessive quantities of these proteins in patients with Marfan

syndrome contributes to the weakening and failure of the aortic wall by reducing the structural integrity of cells in addition to contributing to the absence of network behavior.

TGF-β dysregulation contributes to cardiovascular failure

In healthy individuals, the TGF-β pathway plays a key role in homeostasis and tissue development through regulation of intracellular signalling and tissue repair. TGF-β facilitates the production of contractile and extracellular matrix macromolecules. These molecules include: collagen, fibronectin, thrombospondin, and proteoglycans, which are all critical to the maintenance of healthy aortic tissue.

Typically, FBN-1 regulates the TGF-β pathway through the extracellular microfibrils formed by fibrillin-1 (Groeneveld et al.). The FBN-1 gene directly controls the bioavailability of latent TGF-β within the extracellular matrix. TGF-β bioavailability is its ability to be released and bind to receptors within the extracellular matrix (Y. Li et al.). It is through this bioavailability that TGF-β closely regulates its response to biological stressors. This can be seen especially in the mechanically sensitive aortic region.

Latent TGF-β is stored within the cells to promote targeted distribution of active TGF-B. Disruption in the bioavailability of the TGF-β causes inadequate distribution in the extracellular matrix, which to improper tissue repair, dysregulated production of contractile proteins, and weakened strength of cardiovascular tissues. In patients with Marfan syndrome, mutations in FBN-1 increase the bioavailability of TGF-β, causing excessive tissue growth and weakened connective tissue. Through this causational relationship, mutations in FBN-1 are shown to have a significant role in TGF-β dysfunction (L. Li et al.).

However, the specific effects that FBN-1 mutations have on TGF-β function vary based on the type of mutation: haploinsufficient or dominant negative. To start, patients with dominant negative FBN-1 mutations typically have altered functionality of their fibroblasts. Fibroblasts play a key role in the upregulation of the TGF-β signalling pathway. Normally, fibroblasts regulate tissue repair, maintenance and provide structural support. Mutations in FBN-1 impair the ability of fibroblasts to create a functional extracellular matrix which prevents the production of a robust backbone, and creates a weakened connective tissue. Furthermore, fibroblasts control the bioavailability and receptiveness of TGF-β (Chaudhry et al.). Fibroblasts signal the TGF-β pathway for proteins needed to strengthen or remodel the extracellular matrix (MacFarlane et al.). This happens as a response to mechanical stress or chemical cues. However, in response to the structurally weak fibers produced by the mutated fibrillin-1, the fibroblasts erroneously signal to the TGF-β pathway for the production of more contractile proteins to strengthen the architecture of the extracellular matrix. This creates a feedback cycle of oversignalling, leading to disrupted TGF-β pathways in patients with dominant negative FBN-1 mutations (Sakai et al.).

Similarly, in individuals with haploinsufficient FBN-1 mutation, TGF-β signalling is affected by fibrillin-1 architecture in the extracellular matrix (Wipff et al.). However, unlike dominant negative mutations, the extracellular matrix is disrupted due to a lack of fibrillin production. In haploinsufficient patients, only one working copy of the FBN-1 gene remains, dramatically reducing the amount of fibrillin-1 protein that can be produced. This similarly

contributes to a greater quantity of TGF-β being released and activated in order to compensate for the lack of supporting microfibrils in the extracellular matrix. In the case of haploinsufficient individuals, the lack of microfibrils reduces the ability to store latent complexes, affecting the regulation of the TGF-β pathway (Nistala et al.).

In both variations of FBN-1 mutations, hyperactivation of the TGF-β pathway interferes with normal cellular function through affecting the production of key contractile proteins and extracellular matrix constituents. This directly affects the strength and resilience of cardiac muscle, contributing to potential cardiovascular failure.

mTOR pathway overactivation weakens the aortic wall

The mammalian target of rapamycin (mTOR) signaling pathway is a key regulator of cell growth and metabolism. The pathway creates mTOR, a protein that forms and regulates the mTORC1 protein complex. mTORC1 is highly sensitive to signalling from growth factors such as the TGF-B and mTOR. The protein complex has been shown to positively regulate cell growth and proliferation through supporting anabolic processes such as biosynthesis of proteins, lipids, and organelles. Furthermore, mTORC1 contributes to the synthesis of proteins, such as fibrillin,by increasing mRNA biogenesis and the translation of ribosomal proteins (Laplante and Sabatini).

The mTOR pathway is activated when there is a surplus of growth factors. In patients with Marfan syndrome, the production of TGF-B leads to the overactivation of the mTOR pathway. This is primarily seen in individuals with haploinsufficient FBN-1 mutation, where there is an underexpression of the gene. Underexpression of FBN-1 leads to reduction of fibrillin protein within the body. Fibrillin is a key component to the extracellular matrix, and thus lower quantities constitute a weakened matrix and weakened cellular barrier to insults. The absence of the protein triggers the production of TGF-B (Jiang et al.); a group of signalling molecules that activates the mTOR pathway, thus falsely signalling the production of mTORC1. The presence of mTORC1 promotes the translation of FBN-, increasing fibrillin concentration in the body. However, due to the nature of the FBN-1 mutations, the fibrillin protein is either overexpressed at the translational level or mutated in such a way that prohibits binding to the cell, and leads to a terminally weakened extracellular matrix and thus cellular barrier. This in turn leads to perpetual mTOR signalling, where there is a constant production of mTOR protein and mTORC1 to restore fibrillin to the extracellular matrix. This creates a constant feedback loop between FBN-1, TGF-B, and mTOR, resulting in the chronic hyperactivation of the mTOR pathway often leading to incorporation of weakened or mutated fibrillin proteins into the extracellular matrix. Overall these mutations are involved in the weakening of the cardiovascular tissue and muscles.

Perlecan and Perlecan Function

Perlecan, a component of the extracellular matrix, is a multifunctional heparan sulfate proteoglycan (HSPG). HSPGs are proteins with heparan sulfate chains that play key roles in cell

signaling, development, tissue repair, and filtering. They are typically found in basement membranes, a thin layer of the extracellular matrix. Perlecan acts as an anchor point for fibrillin-1 in basement membranes, thus playing a crucial role in the stability of microfibrils such as actin and collagen (Tiedemann et al.).

Perlecan has been found to co-localize at the interference of elastin (a protein that gives elasticity and recoil to tissues) and fibrillin-1 in the aorta. Fibrillin proteins interact with HSPG by binding with high affinity relative to other proteins. Perlecan additionally has been known to act as a substitute for elastin. Perlecan interacts with extracellular proteins such as fibulin-1, fibronectin, and TGF-B, which influences protein behavior and structure. Perlecan production is upregulated by the TGF-β signaling pathway creating a positive feedback loop in the body between perlecan produced by the FBN-1 mutation and TGF-β activation. excess in turn stimulates growth pathways and influences their bioavailability. This hyper-active state induced by excessive available perlecan directly contributes to an overproduction of proteins like fibronectin which lead to unstable collagen architecture followed by fibril organization disruption.

Fibronectin and integrin

Fibronectin production is increased in individuals with Marfan syndrome. Fibronectin is a protein that is commonly found in the extracellular matrix. Located in submucosal basement membranes and interstitial tissues, fibronectin promotes extracellular matrix adhesion and externally anchors cells to collagens or proteoglycans (Moroni et al.). Fibronectin also plays a crucial role in anchoring organelles within the cytoplasm, helping to provide cellular support to mechanical strain. The adhesion of the extracellular matrix is done through integrins, a group of proteins found on the surface of cells that acts as a receptor. Integrins bind to the extracellular matrix inducing the relay information and influence the cell behavior of surrounding cells or proteins (Alberts et al.).

In normal tissue, fibrillin-1, integrin, and fibronectin work in parody to build the extracellular matrix. Fibrillin-1 proteins support structure by creating longer, more intensive fibrils that are concentrated in areas prone to high mechanical stress (Kinsey et al.). In healthy tissue structure, integrins link the extracellular matrix to the actin cytoskeleton, helping to connect the fibrils to cellular walls. In the aorta, fibronectin is incorporated into actin fibers through integrins, providing contractile support and adhesion to the wall of the vessel.

Dysregulation of fibronectin in patients with Marfan syndrome removes the adhesive ability of the stress fibers. Furthermore, in patients with Marfan syndrome, FBN-1 driven dysregulation of the TGF-β pathway contributes to pathological fibrosis and an overproduction of fibronectin. TGF-β is responsible for the differentiation of fibroblasts, a key producer of fibronectin. In patients with Marfan syndrome, dysregulated TGF-β stimulates higher amounts of fibroblast driven extracellular matrix remodelling. This disrupts the collagen I fibrils, leading to a lower concentration in the extracellular matrix, thus contributing to reduced contractility in the aortic wall (Sapudom et al.).

Abnormal collagen levels contribute to decreased elasticity in aortic wall

Furthermore, in patients with Marfan syndrome, mutated FBN-1 causes hyperactivity of the TGF-β pathway, which contributes to abnormal organization and production of type I, III, and IV collagens. Collagen is a fibrous protein that provides structure and support to tissues that undergo significant quantities of mechanical stress. Typically, collagen type I and type III are found in the heart, and collagen type I, II, and IV are organized into the layers of the aorta (Lindeman et al.). The incorporation of collagen into the cardiac muscle and tissue provide tensile strength and resilience to cyclic deformation–the cardiac strain the heart endures during a cardiac cycle.

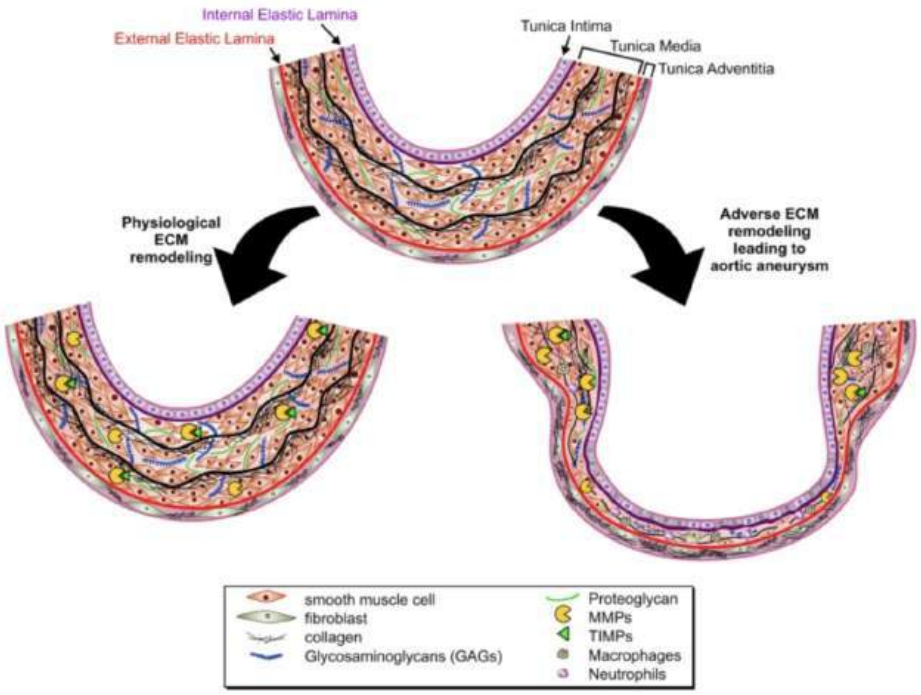


Fig 2. typical organization of aortic cellular components (of note: smooth muscle cells, fibroblasts, collagen), in the tunica intima, media, and adventitia. The concentration of collagen varies between the intima, media, and adventitia, and also demonstrates distinct organization. Also reflected in the figure is the various effects that occur when the cellular components of the aorta are dysregulated or deviate from normal composition (Figure Sourced from: (Jana et al.)

The aortic wall can be divided into three layers: the tunica intima, media, and adventitia (Fig. 2). The tunica intima and media provide crucial support for cardiac mechanical function. The tunica intima and media contain elastic fibers in their wall structures, allowing for proper distention and recoil during the cardiac cycle (de Souza et al.). Contrastingly, the tunica

adventitia serves as an anchor for surrounding tissues and membranes, helping to maintain the elasticity and robustness of the artery.

In the tunica intima, collagen type IV serves as a component of the basal lamina within the basement membrane. The basal lamina is a thin flexible layer of extracellular matrix that binds the endothelial cells to the tunica intima (Urabe et al.). Collagen type IV acts as a scaffold, playing a crucial role in cellular stability and assembly. Collagen is distributed around the elastic fibers and in between vascular smooth muscle cells, providing mechanical support during artery dilation and contraction. A significant reduction in type IV collagen signalling has been observed in patients with Marfan syndrome. This suggests altered architecture of collagen weakens the aortic wall through altered organization in the tunica intima.

Next, the tunica media is characterized by the presence of type I and type III collagen. Typically, collagen type III usually has a more widespread distribution than collagen type I. However, in Marfan syndrome a decrease in collagen type III and an increase in collagen type I has been observed. Although there has been a reduction in collagen type III production, total collagen production still remains higher in Marfan syndrome. Interestingly, the absence of type III collagen has been noticed particularly in areas associated with elastic fiber fragmentation within the aortic tissue. The lack of collagen in these areas constitutes a reduction in the integrity of the elastic fibers in the tunica media. In addition, collagen has been observed to have a greater fibular diameter in the tunica media.

Furthermore, collagen type III fibers are normally located in the spaces between the elastic layers of the aortic wall (intermellar space) and along the elastic layers (elastic lamellae) themselves. This hints towards a reduction of type III collagen signalling within the intermellar space and along the elastic lamellae contours. Given the disrupted distribution and architecture of collagen it is suspected that individuals with Marfan syndrome have a disrupted spatial distribution and uneven weakening of cell walls in the tunica media layer. Lastly, in the tunica adventitia, collagen is normally arranged into a long, ribboning, network of bands that help to provide resilience to cystic deformation. These collagen ribbons allow for flexibility, permitting the arterial wall to distend easily without overstretching.

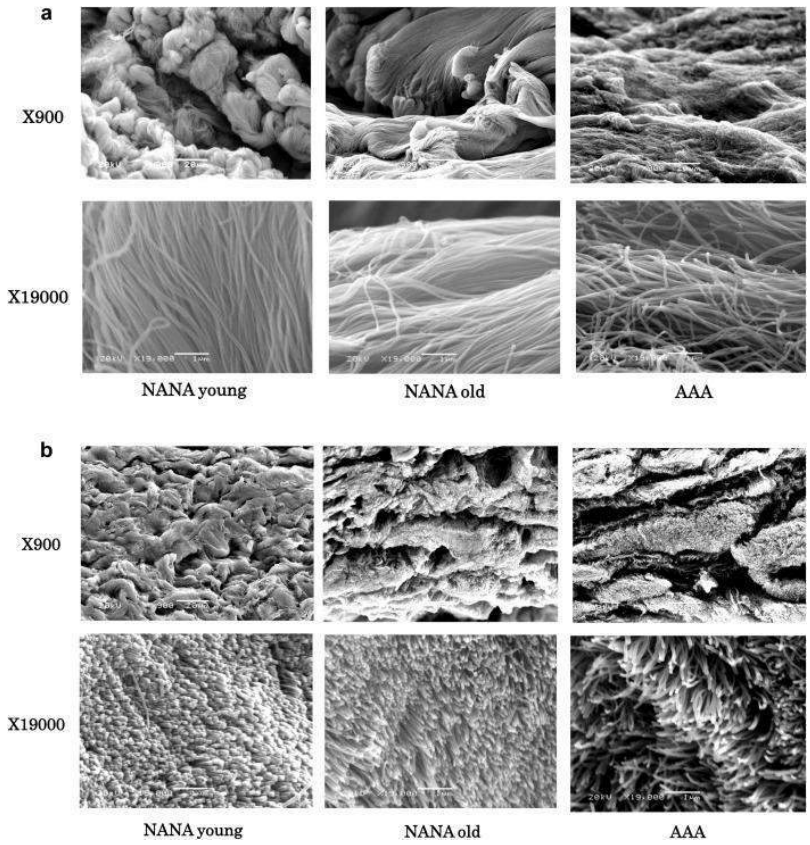


Fig 3: Adventitial tissue fibers in the abdominal aorta. The above figure visualizes collagen fibrils in adventitial tissue through scanning microscopy. Imaging has been done in: A) birds- eye and B) cross sectional views. Collagen fibers were observed in nonatherosclerotic nonanurysmal (NANA) young aortas and in abdominal aortic aneurysm (AAA) aortas. Collagen observed in NANA aortas found fibers to be in rope-like bundles in young aortas, and unraveled and flat in older aortas. Contrastingly, young AAA aortas showed flat and disintegrated collagen strand formation. Collagen also appeared to be 'fluffier' or less tightly wound in AAA. (Figure sourced from:(Urabe et al.).

In Marfan syndrome, there is a complete lack of normal collagen architecture and instead a deposition of thin parallel collagen fibrils (Fig. 3). To start, in contrast to the ribbon-like collagen in a normal aorta, Marfan syndrome collagen runs almost parallel, inhibiting the ability to stretch and thus stiffening the tissue. Furthermore, In healthy aorta, adventitial tissue behaves as a coherent network, dispersing stress equally across the tissue. However, in the adventitial tissue layer in those with Marfan syndrome, the fibers are easily moved when probed, indicating loaded fibrils that are unable to transfer the stress and strain to neighboring tissues. Marfan syndrome cardiovascular tissue shows an absence of network behavior. Collagen in Marfan syndrome has been noted to have a greater fibular diameter in the tunica adventitia. (Roughly corresponds to the thickness of approximately one to two fibrills.) An increase in the size of collagen fibrils further corroborates impaired elasticity and dynamic regulation in the aortic wall of patients with Marfan syndrome.

Actin cycling and cytoskeleton

Actin is a key component to the fibrillar organization in cardiovascular tissue. Actin is a protein that forms filaments, or 'actin fibers', that play crucial roles in cell structural support and movement. In the case of the cardiovascular system, actin fibers help provide integrity to the cardiac muscles, generate force, and muscle contraction. In Marfan syndrome, FBN-1 mutations significantly alter the organization and functionality of the actin cytoskeleton.

Depending on the functionality of the tissue or muscle, actin fibers form a variety of arrangement that provides the backbone for cellular organization. Some common organizations are: antil-parrallel, branch or crosslinked, and parallel bundles. Typically anti-parallel organization is seen in the aorta. Antiparallel is when the actin fibers are organized in opposite directions, creating a denser packing and stronger interaction with motor proteins, thus facilitating strong contraction. A key relationship to note is the interdependence between collagen and actin. Collagen stimulates the polymerization of actin through interactions through cell receptors like integrins.

This relationship is essential for force generation, particularly in the aorta. In individuals with Marfan syndrome, downstream effects of FBN-1 mutations cause changes in the collagen architecture in the aorta and cardiovascular tissue, drastically altering the formation and organization of actin. This in turn creates systematic weakness in tissue and muscle, resulting in further susceptibility to failure under cardiovascular strain.

Discussion

Over the last few decades, there has been a drive to investigate the causational link between cardiovascular and aortic abnormalities in patients with FBN-1 driven Marfan syndrome. It has long been accepted that the mutated FBN-1 gene causes dysregulation of the TGF-β pathway and dysfunction of the fibrillin-1 protein (Dietz).

Through TGF-β pathway oversignalling, FBN-1 mutations have downstream effects on the HSPGs such as perlcan and fibronectin, which in turn causes a perpetual cycle of TGF-β bioactivation. TGF-β stimulates the production of these compounds in the cardiovascular system, which are in turn activators of the growth pathway. This perpetual oversignalling thus constitutes dysfunctional organization of collagen and actin fibers within the aortic wall, inhibiting proper contractile function.

Of note, TGF-B overdrive also activates the mTOR pathway, but this connection can be pharmaceutically treated through drugs such as rapamycin. Rapamycin is known to suppress the mTOR pathway, and is most commonly used for organ transplants. However, in patients with Marfan syndrome the drug has the potential to limit aortic root expansion by mitigating overproduction and stabilizing ECM proteins (Zaradzki et al.). However, further research is needed to determine dosing, and to elucidate and mitigate the side effects.

Additionally, disturbance in the collagen and fibular structure of the aorta have shown to significantly reduce cohesion, preventing the tissue from distributing cardiovascular force evenly. Through disrupted fibular mechanics, the aortic wall is made incredibly susceptible to

overuse and mechanical overload, thus causing characteristic aortic dilations, aneurysms, and dissections.

However, with this in mind it is again important to consider the potential of rapamycin pharmaceuticals. Rapamycin inhibits the mTOR signalling, thus blocking TGF-b from oversignalling collagen production. Pharmaceutically, the drug can be used preventatively to inhibit the overproduction of disrupted collagen, thus having the potential to increase the time before surgical intervention is required.

That being said, further research is needed to fully understand the link between FBN-1 driven phenotypic changes and actin stress fiber organization. Although there is a known link between the two, there is not enough evidence to provide clear reasoning regarding how genotypic variance leads to pathological phenotypes. Actin fibers have the ability to replace mutated elastin and FBN-1. It is important that this link is investigated as understanding the functionality of this replacement can help to develop pharmaceuticals that can activate the actin network, thus increasing the strength of the heart and reducing the need for surgical intervention.

Works Cited

Aalders, Jeffrey, et al. “Effects of Fibrillin Mutations on the Behavior of Heart Muscle Cells in Marfan Syndrome.” Scientific Reports, vol. 10, Oct. 2020, p. 16756. PubMed Central, https://doi.org/10.1038/s41598-020-73802-w.

Alberts, Bruce, et al. “Integrins.” Molecular Biology of the Cell. 4th Edition, Garland Science, 2002. www.ncbi.nlm.nih.gov, https://www.ncbi.nlm.nih.gov/books/NBK26867/.

Chaudhry, Shazia S., et al. “Fibrillin-1 Regulates the Bioavailability of TGFbeta1.” The Journal of Cell Biology, vol. 176, no. 3, Jan. 2007, pp. 355–67. PubMed, https://doi.org/10.1083/jcb.200608167.

Crosas-Molist, Eva, et al. “Vascular Smooth Muscle Cell Phenotypic Changes in Patients With Marfan Syndrome.” Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 35, no. 4, Apr. 2015, pp. 960–72. ahajournals.org (Atypon), https://doi.org/10.1161/ATVBAHA.114.304412.

de Souza, Rodrigo Barbosa, et al. “Exploring Thoracic Aorta ECM Alterations in Marfan Syndrome: Insights into Aorta Wall Structure.” Scientific Reports, vol. 15, no. 1, July 2025, p. 26665. www.nature.com, https://doi.org/10.1038/s41598-025-09665-w.

Dietz, Harry. “FBN1-Related Marfan Syndrome.” GeneReviews®, edited by Margaret P. Adam et al., University of Washington, Seattle, 1993. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK1335/.

Ghoraba, Hashem H., et al. “Biometric Variations in High Myopia Associated with Different Underlying Ocular and Genetic Conditions.” Ophthalmology Science, vol. 3, no. 1, Mar. 2023, p. 100236. PubMed, https://doi.org/10.1016/j.xops.2022.100236.

Groeneveld, Menno Evert, et al. “Betaglycan (TGFBR3) up-Regulation Correlates with Increased TGF-β Signaling in Marfan Patient Fibroblasts in Vitro.” Cardiovascular Pathology, vol. 32, Jan. 2018, pp. 44–49. ScienceDirect, https://doi.org/10.1016/j.carpath.2017.10.003.

Jana, Sayantan, et al. “Extracellular Matrix, Regional Heterogeneity of the Aorta, and Aortic Aneurysm.” Experimental & Molecular Medicine, vol. 51, no. 12, Dec. 2019, pp. 1–15. www.nature.com, https://doi.org/10.1038/s12276-019-0286-3.

Jiang, Yuanyuan, et al. “Marfan Syndrome: Insights from Animal Models.” Frontiers in Genetics, vol. 15, Jan. 2025, p. 1463318. PubMed Central, https://doi.org/10.3389/fgene.2024.1463318.

Kinsey, Rachel, et al. “Fibrillin-1 Microfibril Deposition Is Dependent on Fibronectin Assembly.” Journal of Cell Science, vol. 121, no. Pt 16, Aug. 2008, pp. 2696–704. PubMed, https://doi.org/10.1242/jcs.029819.

Knorst, Ana Paula. “Measuring Mechanical Properties of Aorta Tissue.” Rheolution, 4 Dec. 2025, https://rheolution.com/application-notes/measuring-mechanical-properties-of-aorta-tissue

Laplante, Mathieu, and David M. Sabatini. “mTOR Signaling at a Glance.” Journal of Cell Science, vol. 122, no. 20, Oct. 2009, pp. 3589–94. PubMed Central, https://doi.org/10.1242/jcs.051011.

Li, Li, et al. “The Extracellular Matrix Glycoprotein Fibrillin-1 in Health and Disease.” Frontiers in Cell and Developmental Biology, vol. 11, 2023, p. 1302285. PubMed, https://doi.org/10.3389/fcell.2023.1302285.

Li, Yujia, et al. “Transforming Growth Factor β Latency: A Mechanism of Cytokine Storage and Signalling Regulation in Liver Homeostasis and Disease.” JHEP Reports, vol. 4, no. 2, Feb. 2022, p. 100397. ScienceDirect, https://doi.org/10.1016/j.jhepr.2021.100397.

Lindeman, Jan H. N., et al. “Distinct Defects in Collagen Microarchitecture Underlie Vessel-Wall Failure in Advanced Abdominal Aneurysms and Aneurysms in Marfan Syndrome.” Proceedings of the National Academy of Sciences of the United States of America, vol. 107, no. 2, Jan. 2010, pp. 862–65. PubMed, https://doi.org/10.1073/pnas.0910312107.

Loeys, Bart L., et al. “The Revised Ghent Nosology for the Marfan Syndrome.” Journal of Medical Genetics, vol. 47, no. 7, July 2010, pp. 476–85. PubMed, https://doi.org/10.1136/jmg.2009.072785.

MacFarlane, Elena Gallo, et al. “TGF-β Family Signaling in Connective Tissue and Skeletal Diseases.” Cold Spring Harbor Perspectives in Biology, vol. 9, no. 11, Nov. 2017, p. a022269. PubMed, https://doi.org/10.1101/cshperspect.a022269.

Miyakawa, Ayumi A., et al. “RAPAMYCIN ACTIVATES TGF RECEPTOR INDEPENDENTLY OF ITS LIGAND: IMPLICATIONS FOR ENDOTHELIAL DYSFUNCTION.” Clinical Science (London, England : 1979), vol. 132, no. 4, Feb. 2018, pp. 437–47. PubMed Central, https://doi.org/10.1042/CS20171457.

Moroni, Lorenzo, et al. “Tissue Engineering: An Introduction.” Tissue Engineering, Elsevier, 2015. research.utwente.nl, https://doi.org/10.1016/B978-0-12-420145-3.00001-8.

Nistala, Harikiran, et al. “Fibrillin-1 and -2 Differentially Modulate Endogenous TGF-β and BMP Bioavailability during Bone Formation.” Journal of Cell Biology, vol. 190, no. 6, Sept. 2010, pp. 1107–21. Silverchair, https://doi.org/10.1083/jcb.201003089.

Sakai, Lynn Y., et al. “FBN1: The Disease-Causing Gene for Marfan Syndrome and Other Genetic Disorders.” Gene, vol. 591, no. 1, Oct. 2016, pp. 279–91. PubMed, https://doi.org/10.1016/j.gene.2016.07.033.

Salik, Irim, and Prashanth Rawla. “Marfan Syndrome.” StatPearls, StatPearls Publishing, 2026. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK537339/.

Sapudom, Jiranuwat, et al. “The Interplay of Fibronectin Functionalization and TGF-B1 Presence on Fibroblast Proliferation, Differentiation and Migration in 3D Matrices.” Biomaterials Science, vol. 3, no. 9, Sept. 2015, pp. 1291–301. Silverchair, https://doi.org/10.1039/c5bm00140d.

Silverman, D. I., et al. “Life Expectancy in the Marfan Syndrome.” The American Journal of Cardiology, vol. 75, no. 2, Jan. 1995, pp. 157–60. PubMed, https://doi.org/10.1016/s0002-9149(00)80066-1.

Takeda, Norifumi, et al. “Impact of Pathogenic FBN1 Variant Types on the Progression of Aortic Disease in Patients With Marfan Syndrome.” Circulation: Genomic and Precision Medicine, vol. 11, no. 6, June 2018, p. e002058. ahajournals.org (Atypon), https://doi.org/10.1161/CIRCGEN.117.002058.

Takeda, Norifumi, and Issei Komuro. “Genetic Basis of Hereditary Thoracic Aortic Aneurysms and Dissections.” Journal of Cardiology, vol. 74, no. 2, Aug. 2019, pp. 136–43. PubMed, https://doi.org/10.1016/j.jjcc.2019.03.014.

Tiedemann, Kerstin, et al. “Microfibrils at Basement Membrane Zones Interact with Perlecan via Fibrillin-1 *.” Journal of Biological Chemistry, vol. 280, no. 12, Mar. 2005, pp. 11404–12. www.jbc.org, https://doi.org/10.1074/jbc.M409882200.

Urabe, Go, et al. “Structural Analysis of Adventitial Collagen to Feature Aging and Aneurysm Formation in Human Aorta.” Journal of Vascular Surgery, vol. 63, no. 5, May 2016, pp. 1341–50. ScienceDirect, https://doi.org/10.1016/j.jvs.2014.12.057.

Wipff, Julien, et al. “[Interactions between fibrillin-1 and tgf-beta: consequences and human pathology].” Medecine Sciences: M/S, vol. 25, no. 2, Feb. 2009, pp. 161–67. PubMed, https://doi.org/10.1051/medsci/2009252161.

Wozniak-Mielczarek, L., et al. “Differences in Cardiovascular Manifestation of Marfan Syndrome Between Children and Adults.” Pediatric Cardiology, vol. 40, no. 2, 2019, pp. 393–403. PubMed Central, https://doi.org/10.1007/s00246-018-2025-2.

Zaradzki, M., et al. “Short-Term Rapamycin Treatment Increases Life Span and Attenuates Aortic Aneurysm in a Murine Model of Marfan-Syndrome.” Biochemical Pharmacology, vol. 205, Nov. 2022, p. 115280. PubMed, https://doi.org/10.1016/j.bcp.2022.115280.

Zeigler, Sanford M., et al. “Pathophysiology and Pathogenesis of Marfan Syndrome.” Advances in Experimental Medicine and Biology, vol. 1348, 2021, pp. 185–206. PubMed, https://doi.org/10.1007/978-3-030-80614-9_8.

The Role of LSM14B in Eukaryote Oocyte Development and Cancer

By Anvita Lakhotia

1. Abstract

In the final stages of oocyte development, transcription stops and resumes only after fertilization of the oocyte. Thus, the stored mRNAs are essential to oocyte growth as there are no other mRNAs available to synthesize new proteins. As a result, the storage and maintenance of mRNAs is crucial during oocyte development; however, how oocytes do so has been largely unknown. Recent studies suggest that RNA Binding Proteins (RBPs) play critical roles in this process. One such RBP is LSM14B, a protein functional in early zebrafish and mammalian oocyte development. In zebrafish, LSM14B seems to play an important role in fertility and regulating the length of mRNA poly-A tails to prevent mRNA degradation. In mammals, LSM14B regulates oocyte meiotic maturation- especially in the transition to Anaphase 1- and female fertility in mice. LSM14B plays important roles in cancer development as well, specifically Colorectal and Breast cancer, and Hepatocellular Carcinoma. This review covers recent findings and efforts in the research field of RBP (LSM14B)-mediated oocyte development, and further its potential functions in other cellular and biological contexts.

Keywords gene, LSM14B, mRNA, zebrafish, oocyte maturation, fertility, cancer

2. Introduction

RNA binding proteins are proteins bound to mRNAs and help facilitate many different functions, including: intracellular transport, entry to and exit from translational activity, and recruitment of or protection from the cell's degradation machinery- such as autophagy. [27] There are over 1000 RBP encoding genes, and hundreds of RBPs. Each RBP has a distinct RNA binding specificity, and protein-protein interaction domains that mediate the unique assembly of RBPs on every RNA, thereby sculpting the RNA for processing and function by cellular machinery. RNA binding proteins hold varied roles in different processes, including the post-transcriptional regulation of gene expression, cell differentiation, metabolism, and oocyte development. Because of the diversity of RBPs, cells can regulate alternative splicing of genes and respond to environmental changes. Without RNA binding proteins, RNA quickly decays, leading to disarray in the production of specific gene products. [25]

One of the RBPs that plays a role in mRNA maintenance is LSM14B. LSM14B is a part of the LSM14 protein family, a highly conserved protein family implicated in the regulation of maternal mRNA metabolism and oogenesis in several species. Further, recent studies suggest LSM14B is also expressed in several cancers and plays critical roles. Aberrant overexpression of LSM14B has also been identified in various cancers, including colorectal cancer (CRC), hepatocellular carcinoma (HCC), and breast cancer. Although the specific mechanism of LSM14B function has been not entirely understood, growing evidence suggests this gene has a pivotal role in the context of embryonic development and disease. [1,2]

Given the growing body of research spanning structural molecular biology, developmental genetics, and clinical oncology, a comprehensive synthesis of LSM14B function is needed. This review aims to examine the multifaceted functions of LSM14B. It outlines the structural and molecular features of the LSM14 family, summarizes its necessity in eukaryotic oocyte maturation and maternal mRNA dynamics, and discusses emerging findings linking LSM14B dysregulation to cancer progression. Finally, it highlights current limitations in the field and suggests future directions for therapeutic targeting and biological investigation.

3. Molecular characteristics of LSM14 Family

LSM14B was originally identified in Xenopus Laevis and the P-body of human cells and is implicated in linking oocyte maturation to female fertility through regulation of the P-body [2]. In 2023, Wan et al. performed quantitative PCR and immunoblot assays in mouse oocytes and found that it is highly expressed during oocyte development. The expression level decreases during meiotic maturation, becoming negligible in two-cell embryos (which develop after fertilization). Researchers found "LSM14B had highest localization at mitochondria in meiosis I stage in oocytes and appeared to be involved in formation of the mitochondria-associated ribonucleoprotein domain (MARDO), a mitochondria-associated membraneless compartment that stores maternal mRNAs in oocytes."[3] This prompted researchers to experiment more with the gene to better understand its role in oocyte maturation.

First, looking at the molecular characteristics of the LSM14 gene family, the researchers first produced pieces of human proteins to study how they interact. They produced the LSM domain of LSM14 (the first 80 amino acids), pieces of 4E-T, and the C-terminal RecA domain of DDX6. They also produced variants of the proteins with mutations. Next, they used X-Ray Crystallography - in which they grew crystals of protein complexes and used X-ray diffraction-to observe how the proteins attach to each other. Crystallography allowed them to map out physical contact surfaces and binding orientation between these proteins in a three dimensional molecular structure. Their results showed how 4E-T wraps around LSM14, and how short motifs in LSM14 bind to DDX-6. [4,16,20]

To quantify how the proteins interacted and identify specific contact points, the researchers used a number of methods. These included pull-down assays, mutagenesis, and Isothermal Titration Calorimetry (ITC). The pull down assays were used to detect whether or not the proteins even interacted. Mutagenesis helped identify specific amino acids integral to protein interaction, and ITC measured the degree of binding- from loose to tight- and gave information on whether the interactions were strong or weak. In the crystal of the proteins, 4E-T wraps around the LSM14 domain and makes contact through hydrophobic interactions. Specifically, there are also several pockets on the LSM14 surface where conserved amino acids from 4E-T insert like pegs into holes in the LSM14 domain. The pull down assays further confirmed these interactions. Mutagenesis of key hydrophobic amino acids in LSM14 or in 4E-T led to weakened binding, further confirming the importance of hydrophobic interaction in the binding of these proteins. To continue dissecting the nature of the binding between the LSM14 domain and 4E-T, researchers used a luciferase reporter assay, which measures repression of an mRNA linked to a

glow-in-the-dark enzyme. They found that the LSM domain of LSM14 can repress translation of a tethered mRNA, but is unable to do so without binding 4E-T. LSM14's N-terminal LSM domain specifically binds the C-terminal region of 4E-T with hydrophobic interactions. This interaction is important for silencing target mRNAs, an important function of LSM14B in oocyte maturation. [4,19,20]

LSM14 binds DDX-6 differently to 4E-T: through the use of two motifs. Previous studies have shown LSM14 has three conserved motifs near its middle: FDF, FFD, and TFG. Previous research has also revealed the FDF motif can bind DDX6, but the roles of the other two motifs were still unclear. The crystal structures of both proteins reveal that LSM14 binds the C-terminal of DDX6 in a bipartite interaction: FDF and TFG are spaced far apart and both contact DDX6 in distinct spots. Mutagenesis of the motifs confirms their uses, as mutating key phenylalanines in FDF or TFG reduced binding to DDX6. The interaction between LSM14 and DDX6 is useful because DDX6 plays an important role in mRNA gene silencing by repressing translation and enhancing mRNA decapping and decay. [4,20]

Although most research to date has been on the molecular structure of LSM14 genes rather than specifically LSM14B, much of the research still serves to inform us of the current structures and uses of LSM14B in mRNA silencing.

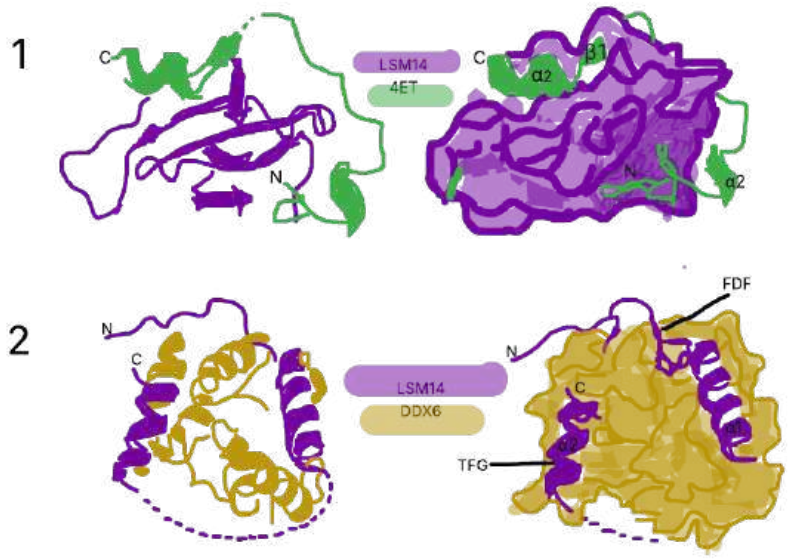


Figure 1: Crystal structure of the N-terminal LSM domain of human LSM14 in complex with a conserved C-terminal 4E-T fragment. The diagram depicts the likely solution structure obtained by combining two fragments of 4E-TC that undergo crystallization-induced domain swapping Left: Cartoon representation of the complex. Right: LSM domain of LSM14 shown in surface representation and 4E-TC residues critical for complex formation are shown in stick format. Secondary structure elements are not shown.

Figure 2: Left: Crystal structure of the DDX6C in complex with a LSM14 fragment containing the FDF, FFD, and TFG motifs. The FFD motif is not labeled, but is denoted with a blue dashed line. Secondary structure elements are not shown and involved motifs are labeled. Right: Surface representation of DDX6C. The bound LSM14 peptide is shown as a cartoon representation, and residues from LSM14 mediating interaction with DDX6C are shown in stick format.

4. Functional roles of LSM14B in zebrafish oocyte development

First, the experimenters completed immunofluorescence of LSM14B and Vasa, a germline marker. It showed that LSM14B was enhanced in early stage oocytes. These results confirmed that LSM14B is specifically expressed in early stage zebrafish oocytes. [1]

Two mutant versions of LSM14B were created by the CRISPR/Cas 9 system. By completing polymerase chain reaction tests, it was confirmed there was a deficiency of LSM14B in the mutants. This, in turn, confirmed nonsense RNA mutations led to a degradation of LSM14B in the mutant alleles. The researchers also found no significant phenotypic differences between the two mutants, so they chose to move forward in experimentation with one mutant: LSM14B^+11/+11 (which will now be denoted in this paper as M). [1]

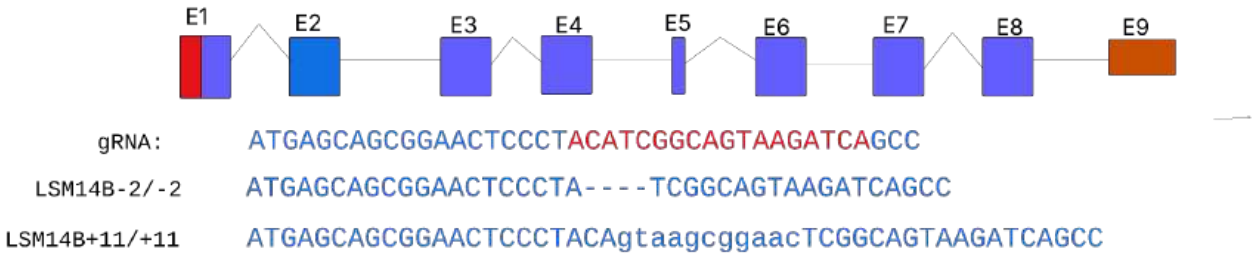


Figure 3: Schematic diagram of the zebrafish LSM14B gene structure and location of gRNA target sites from E1 only. The gRNA target sites are highlighted in red. PIcture depicts the original LSM14B gene vs the two mutants developed by researchers. They chose to move forward with LSM14B^+11/+11 (M). [1]

To further observe the implications of a mutant LSM14B in zebrafish oocyte development, the researchers mated wild type males with M females in one trial, and mated Wild Type (WT) females with M males in another trial. They found that M male mutants exhibited typical fertility at four months post fertilization, similar to the wild type males. By contrast, the female LSM14B- mutants were unable to spawn. To further analyze infertility, the researchers closely examined the oocyte development of the LSM14B female mutants. Fertility tests were conducted by mating 4 month old M males and females with 4 month old fertile WT females and males. Zebrafish complete their development and reach sexual maturity within 3-4 months. [1,20] Individuals unable to spawn after at least 10 attempts were labeled infertile. They found that M females differed significantly to WT females in ovary morphology. Specifically, the gonadosomatic index (GSI) displayed the ovary of the mutant females was placed significantly lower than those of the wild type females. Additionally, the ovary of the wild type females contained all stages of follicles, including the PG, PV, early vitellogenic (EV), mid-vitellogenic (MV), late vitellogenic (LV), and full-grown (FG) pre-ovulatory stages. However, oocytes in the M females were arrested at the PG stage, only a few PV, and no EV, MV or FG stage oocytes were found. These results show that LSM14b plays an important role in the oocyte PG - PV transition in zebrafish, suggesting that LSM14B plays a role in fertility of zebrafish. [1]

Next, the researchers chose to focus on poly-A tails. LSM14B keeps mRNA from degrading during the translation-inhibited state of oocyte development (i.e when RNA is not translated into proteins). Poly A tails, along with the 5' cap, both prevent mRNA degradation as it leaves the nucleus and travels to the ribosome. The similar properties between LSM14B and poly-A tails encouraged researchers to experiment on poly-A tails. They extracted total RNAs from the WT and M mutants at 4 months. It was found that LSM14B regulates the length of the mRNA poly-A tail in early zebrafish oocytes, thus maintaining the translation-inhibited state of mRNA during early oocyte development. [1,17,18]

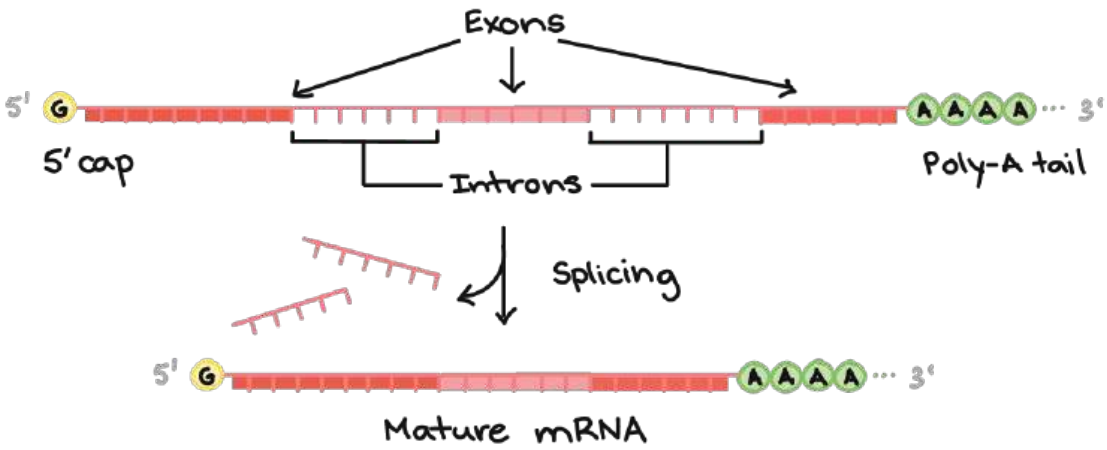


Figure 4: Diagram of typical mRNA strand. 5' cap and Poly-A tails are both integral parts of the mRNA strand. When Poly-A tails are removed or shortened- known as deadenylation-, it releases PABPC (a protein) and leads to mRNA degradation. The length of poly A tails thus must be regulated by gene LSM14B. [17] Figure taken from [26] No changes were made to the original figure.

5. Functional roles of LSM14B in mammalian oocyte development

LSM14B also plays an important role in the development of other organisms, such as mammals. As both LSM14A and LSM14B have been found to play a large role in the regulation of maternal mRNA metabolism, they were chosen to further investigate the potential participation of newly identified oocyte RBPs in maternal mRNA regulation. The expression of the gene itself was first confirmed using the Western blot and Immunofluorescence analyses. Western blotting provides validation that there is an antibody, but is typically performed using denaturation. On the other hand, Immunofluorescence detects proteins in their native state. It was found that LSM14B was exclusively expressed in the oocytes at all follicle stages, and not in granulosa cells or any of the other tissues. As granulosa cells develop much later in the organism than follicle cells do, the results suggested that LSM14B might function as an oocyte‐specific RBP in mammals. [3, 21]

The role of LSM14B in early oocyte development is further investigated using mice carrying a targeted knockout allele of LSM14B (LSM14B ^tm1a(KOMP)Mbp) induced by

reporter-tagged insertion- when a reporter gene is attached to a gene of interest to keep track of the installing [28] When LSM14b ^tm1a(KOMP)Mbp (now shortened to KO) was expressed homozygously in females, LSM14B protein production was undetected in the oocyte. This shows that KO is a null allele. While the KO homozygous male mice showed normal fertility, the KO females were found to be infertile, and ovulated fewer eggs than the wild type (WT) females. Additionally, when the eggs of the KO females were fertilized by WT sperm, the oocyte gradually degenerated in culture. More specifically, 80% of the LSM14B-KO ovulated eggs were arrested at interphase, as opposed to the WT eggs that arrested at metaphase II. Therefore, it can be concluded that LSM14B is needed for oocyte meiotic maturation and female fertility in mice. [5]

To find more detail about the arrests in meiotic progression, oocytes were injected with green fluorescent protein (GFP) to tag tubulin, and mCherry to tag histone (both meiotic components). It was found that the MI-to-anaphase I (AI) transition and the first polar body emission (PBE) were both significantly delayed by three to five hours. Additionally, along with the delayed Anaphase I transition, about 73% of KO oocytes were shown to have prolonged activity at their kinetochores, displaying prolonged activity of the spindle checkpoint. To further investigate the arrest at the anaphase phase, the activity of the Anaphase-promoting complex (APC) was estimated by injecting WT and KO oocytes with mRNAs encoding Venus-tagged CCNB1 (Cyclin B1) and mCherry-tagged PTTG1 (Securin) and observed changes in oocytes via live-cell imaging [5].

To better understand the mechanical causes of how KO causes oocyte meiotic defects, researchers investigated transcriptome- and proteome-wide changes in gene expression in LSM14B-KO oocytes. Researchers found the mis-expression of 2285 genes using transcriptomic analysis, leading them to hypothesize that the upregulation and downregulation of certain genes could account for the phenotypes of prolonged meiosis and delayed transition to Anaphase I. [5]

6 The Role of LSM14B in Cancer Development

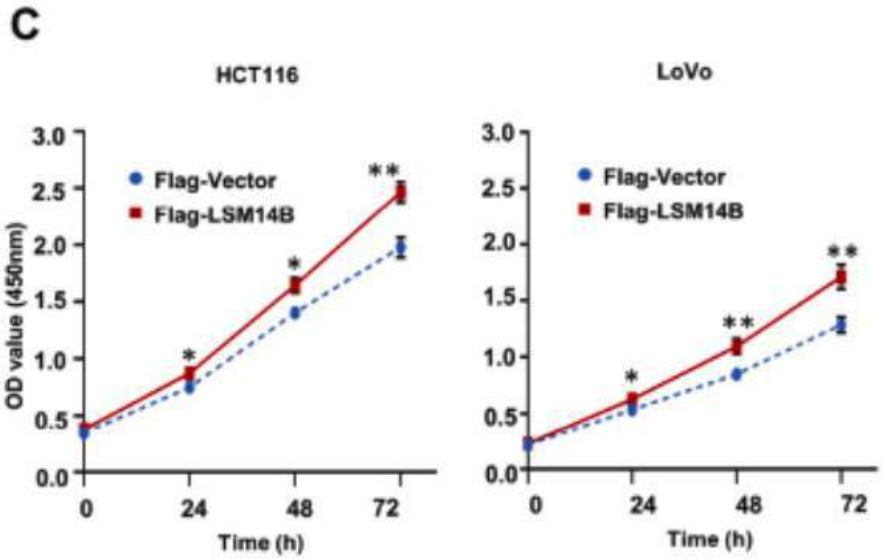
Although LSM14B has been found to play a major role in female fertility, the gene and the LSM gene family has also been thoroughly researched in cancer development. Because the LSM14 gene family regulates mRNA, the genes play a large role in cancer development, especially increasing production of growth-promoting proteins, avoiding cell death, adapting to stress, and evading immune response. Higher LSM14B expression, specifically, is found to play a role in colorectal cancer, breast cancer, and hepatocellular carcinoma. [6,7,10]

6.1 LSM14B in Colorectal Cancer

This study included nine pathologically confirmed CRC patients who underwent surgical resection at Peking University Cancer Hospital & Institute (Beijing, China) between September 2009 and October 2011. From each patient, tumor tissues and historically normal adjacent tissues were compared. The researchers conducted real-time polymerase chain reaction (PCR) tests on all 18 samples. First, they used reverse transcriptase to convert RNA into complementary DNA

(cDNA) in order to run a PCR test on it. Next, they used green fluorescent protein (GFP) to detect DNA amplification. They tagged the DNA by cloning the GFP in frame with the target protein at either the N- or C-terminus of the amino acid chain. [6, 23] Then, the researchers designed specific primers for each gene being tested: the primers for LSM14B were 5′-GAAGACCGTCCCACAGATAGG-3′ (forward) and 5′-GCTGAGCTTTCGGAGGTTCA-3′ (reverse). All of the components of the experiment came together as the researchers calculated the fold change in gene expression to find how much more or less the genes were expressed in the tumor tissue cells as opposed to the normal tissue cells. The results confirmed that LSM14B is overexpressed in CRC cells. [6]

The researchers then further delved into the role of LSM14B in CRC cells. They transfected (infected a cell with a full-length free nucleic acid) LSM14B plasmid into HCT116 and LoVo cells, which increased LSM14B mRNA levels 75.59-fold and 59.02-fold, respectively. They used the HCT116 and LoVo cells as these cells already show relatively low levels of LSM14B. The growth of the CRC cells was determined by CCK8 assays. Colony formation was assessed by seeding CRC cells into a 6-well plate and incubating them for 14 days. Cell colonies were then fixed with paraformaldehyde, stained violet, and counted. Finally, transwell chambers were used to observe migration and invasion. Cells were seeded into the upper chamber in a serum-free medium, and the bottom chamber was coated with 15% Fetal Bovine Serum (which acts as a chemical attractant). Additionally, the upper chamber was coated with Matrigel for the invasion assay, which mimics extracellular tissue. Results displayed that overexpression of LSM14B caused colony formation in HCT116 and LoVo cells, and also promoted the invasion and migration of the same cells.These properties are in line with those of cancerous cells, showing that LSM14B helps CRC cells grow faster and invade surrounding tissue- thus displaying LSM14B is an oncogenic gene in colorectal cancer. [6,24]



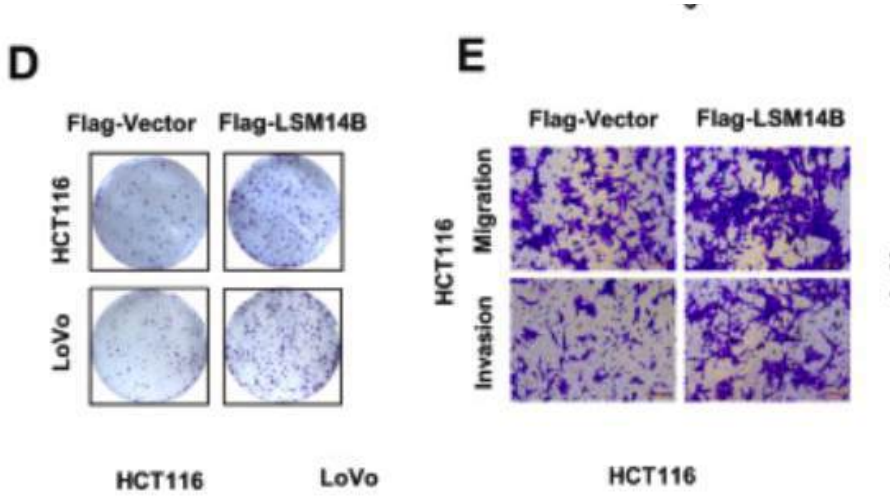


Figure 5: (C) LSM14B overexpression promoted the proliferation of HCT116 and LoVo cells. (D) Overexpression of LSM14B promoted colony formation in HCT116 and LoVo cells. The bar graph shows the number of cells. (E) Overexpression of LSM14B enhanced the migration and invasion of HCT116 and LoVo cells. Bar charts indicate numbers of migrated cells. Figure taken from [6]

6.2 LSM14B in Breast Cancer

Research has also found that the gene LSM14B plays a role in breast cancer. In one study, 3593 patients with breast cancer and their mRNA expression values were utilized to investigate the clinical relevance of LSM family genes in cancer, using bioinformatics and publicly available databases. First, the UALCAN database was used to compare mRNA levels of LSM14B in normal breast tissue samples (114) vs breast tumors (1097). The MethSurv database was used to analyze methylation of CpG sites in LSm14B, as 70-80% of these sites are methylated in mammals [7, 29]. The high mutagenic tendencies of these sites help to explore epigenetic regulation in tumors. Next, the METABRIC and TCGA datasets identified genes co-expressed with LSM14B, and the GO and KEGG pathways were analyzed using Cytoscape and ClueGo to determine biological/cellular components and molecular functions associated with LSM14B. Next, the Kaplan-Meir Plotter assessed correlations between LSM14B expression and in BRCA patients. Timer 2.0 evaluated correlations between LSM14B mRNA levels and tumor-infiltrating immune cells across BRCA subtypes. Finally, the Genomic Data Commons was used to compare LSM14B expression in normal vs tumor tissue and identified correlated genes. Results showed LSM14B is overexpressed in breast tumor tissue, and expression tends to increase in higher tumor stages. Mutation/alteration rates of LSM14B reach 23% in TCGA BRCA patients- one of the highest rates among other LSM gene family members. Next, high LSM14B expression is associated with poor recurrence free survival-HR [Hazard Ratio] = 1.45, 95% CI [Confidence Interval]: 1.23–1.7, $p = 6 \times 10^{-6}$. Finally, LSM14B expression positively correlates with CD4+ T-cell markers, with a correlation coefficient of 0.095. [7]

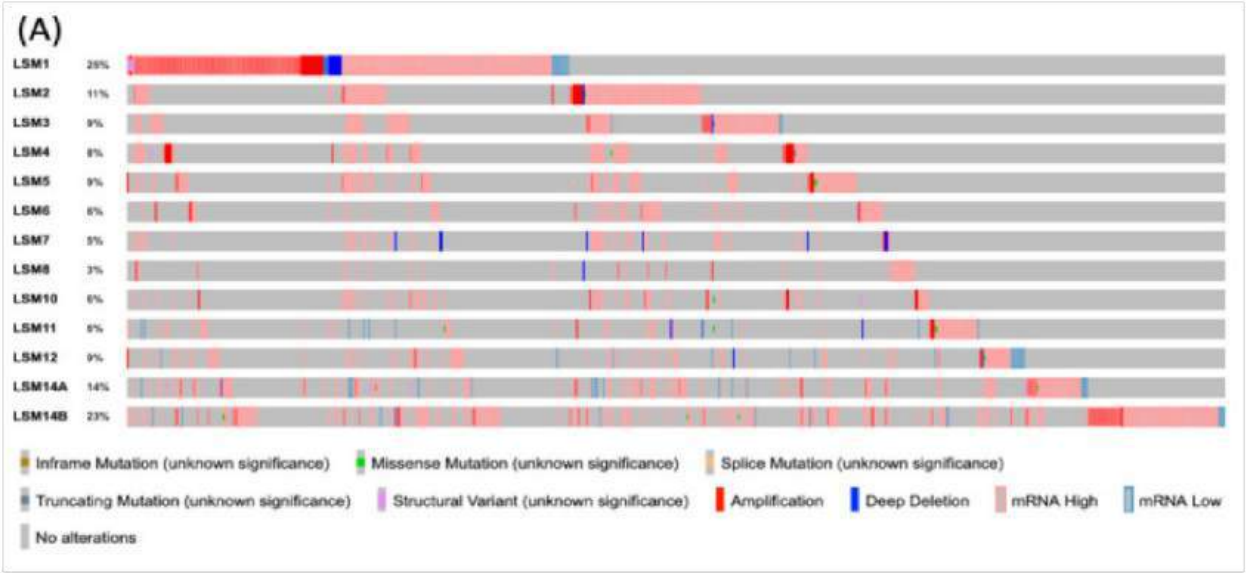


Figure 6: Summary of alterations in LSMs in TCGA PanCancer Atlas (n = 1082). The red bars show gene amplification, blue bars stand for deep deletions, green bars are missense mutations. This plot shows a high alteration rates of LSM1 (25%), LSM2 (11%) LSM4 (8%), LSM14B (23%). Figure taken from [7]

6.3 LSM14B in Hepatocellular Carcinoma

Finally, research has identified LSM14 family members as unfavorable biomarkers in hepatocellular carcinoma through multiple bioinformatics methods and *in vitro* studies. First, the expression differences of LSM family members in normal tissues vs HCC (hepatocellular carcinoma) tissues was evaluated. They found LSM family members were all overexpressed compared to normal tissues. [11] Next, the researchers used a Kaplan-Meier survival analysis to better understand the prognostic value (the overall and disease-free survival) of LSM family members in HCC. [11,12] They also incorporated a LASSO regression analysis to narrow down the most critical prognosis-associated LSM genes. The final risk score model they created focused on four specific genes: LSM5, LSM10, LSM12, and LSM14B. Patients sorted into the "high risk" group had significantly shorter survival times. [11] Research next dove deeper into the biological mechanisms driven by the LSM genes, using a Gene Set Variation analysis. They checked for Pearson correlations between LSM expressions and 13 oncogenic pathways. Results showed that LSM family members were involved in cell cycle processes (including repair, cycle checkpoints, and mitotic spindles). Specifically, three genes- LSM12, LSM14A, and LSM14B- showed powerful connections to the PI3K-AKT-mTOR signaling network and T-cell receptor pathways. The PI3K/AKT/mTOR (PAM) signaling pathway is a highly conserved signal transduction network in eukaryotic cells that promotes cell survival, cell growth, and cell cycle progression. The T-cell receptor pathways promote a number of signaling cascades that dictate cell fate through regulating cytokine production, cell survival, proliferation, and differentiation. The activity of the genes in these two pathways suggest a role in both tumor growth and

immunity. [11,13,14] To expand upon the role in immunity, the study next used single-sample gene set enrichment analysis (ssGSEA) and the ESTIMATE algorithm. The researchers calculated immune scores and evaluated the infiltration levels of various immune cells (like CD8+ T cells and Th2 helper cells). They also tracked the correlation between LSM expression and known immune checkpoint markers (like PD-L1, B7-H3, and PVR). The experiment concluded that tumors expressing LSM12, LSM14A, and LSM14B had an immunosuppressive environment, which shields cancers against novel immunotherapies with a mix of cellular and soluble components that prompt tumor immunity and progression. [11,15] Finally, to see the data unfold in real time, the researchers performed wet-lab experiments. They used laboratory cell lines to perform knockdown experiments (using siRNA) to silence the expression of the target genes—LSM12, LSM14A, and LSM14B. They then conducted cellular assays to evaluate changes in cell proliferation and invasion. Knocking down LSM12, LSM14A, or LSM14B significantly halted the growth of HCC cells and dramatically impaired their ability to invade and spread. This validated that these specific LSM proteins physically drive the growth and immunity of liver cancer cells. [11]

7. Conclusion, Limitations, and Future Directions

LSM14B, as a RNA binding protein, plays an important role in the regulation of maternal mRNA metabolism and oogenesis. The LSM14 domain attaches to proteins 4E-T and DDx6, through hydrophobic interactions and short motifs, respectively, in order to silence target mRNAs. This allows oocyte maturation to proceed by repressing translation in mammals. LSM14B-knockout mouse oocytes exhibit severe meiotic defects, characterized by delayed Anaphase-I transitions, prolonged spindle assembly checkpoint activity, and eventual degeneration. Similarly, in zebrafish, LSM14B is highly expressed in early stage oocytes to regulate poly-A tails and keep mRNA from degrading. Loss of LSM14B disrupts mRNA poly-A tail maintenance, causing a failure in the folliculogenesis transition, and leading to female infertility. Without LSM14B, reproduction of zebrafish and many mammalian species would be virtually impossible, explaining the high conservation of the gene and the LSM14 family.

Beyond its necessity in fertility and oocyte maturation, dysregulated LSM14B expression heavily influences cancer pathogenesis. In Colorectal Cancer (CRC), overexpression directly drives cell proliferation, colony formation, tissue migration, and invasion. In Breast Cancer (BRCA), elevated LSM14B levels correlate with advanced clinical stages, high mutation/alteration frequencies (reaching up to 23%), and significantly poorer recurrence-free survival outcomes. Finally, in Hepatocellular Carcinoma (HCC), LSM14B works alongside related family members (e.g., LSM12 and LSM14A) to drive oncogenesis through the PI3K-AKT-mTOR signaling network, create an immunosuppressive tumor microenvironment, and promote malignant cell growth and invasion.

This review was limited due to the fact that it relied on finite research on the gene LSM14B. Expanded research on this gene should focus on targeted cancer therapies, to better understand how small-molecule inhibitors or targeted RNA therapies could disrupt

LSM14B-mediated oncogenic pathways (such as PI3K-AKT-mTOR). Additionally, further research can clarify the precise molecular cascades through which LSM14B regulates poly-A tail lengths and interacts with membraneless organelles like the MARDO during oocyte maturation. A new research opportunity is also exploring whether targeted dietary or metabolic interventions can modulate LSM14B expression to protect against or reverse obesity-induced oocyte quality deterioration. Ultimately, further unraveling the dual role of LSM14B across developmental and pathological conditions will open promising avenues for both reproductive medicine and targeted cancer therapy.

8. Acknowledgements

I would like to acknowledge the CCIR (Cambridge Centre for International Research) for the opportunity to enroll in their research program. Furthermore, I would like to thank my academic mentor, Dr. Mamiko Yajima, for her guidance throughout the research process.

Works Cited

[1] Acharjee, Santosh, et al. "Mechanisms of DNA methylation and histone modifications." Progress in Molecular Biology and Translational Science, vol. 197, 2023, pp. 51–92, https://doi.org/10.1016/bs.pmbts.2023.01.001.

[2] Bose, Priyom. "Luciferase: A Powerful Bioluminescent Research Tool | The Scientist." TheScientist, July 2024, www.the-scientist.com/luciferase-a-powerful-bioluminescent-research-tool-72013.

[3] Brandmann, Tobias, et al. "Molecular Architecture of LSM14 Interactions Involved in the Assembly of mRNA Silencing Complexes." PubMed Central, 6 Mar. 2018, pmc.ncbi.nlm.nih.gov/articles/PMC5881628/#embj201797869-abs-0001.

[4] Carter, Matt, and Jennifer Shieh. "Making and using transgenic organisms." Guide to Research Techniques in Neuroscience, 2015, pp. 253–271, https://doi.org/10.1016/b978-0-12-800511-8.00012-5.

[5] Clark, David P., and Nanette J. Pazdernik. "Protein structure and function." Molecular Biology, 2013, https://doi.org/10.1016/b978-0-12-378594-7.00048-2.

[6] "Eukaryotic Pre-Mrna Processing | RNA Splicing (Article)." Khan Academy, www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/transcription-and-rna-processing/a/eukaryotic-pre-mrna-processing. Accessed 22 July 2026.

[7] Farrar, Shelley. "GFP-Tagging in Fluorescence Microscopy." News Medical, 2019, www.news-medical.net/life-sciences/GFP-tagging-in-Fluorescence-Microscopy.aspx.

[8] Ghosh, Sankar. "T Cell Receptor Signaling | Cell Signaling Technology." Cell Signaling Technology, 2014, www.cellsignal.com/pathways/t-cell-receptor-signaling.

[9] Glaviano, Antonino, et al. "PI3K/AKT/mtor signaling transduction pathway and targeted therapies in cancer." Molecular Cancer, vol. 22, no. 1, 18 Aug. 2023, https://doi.org/10.1186/s12943-023-01827-6.

[10] Kishore, Jugal, et al. "Understanding survival analysis: Kaplan-Meier estimate." International Journal of Ayurveda Research, vol. 1, no. 4, 2010, p. 274, https://doi.org/10.4103/0974-7788.76794.

[11] Li, E., and Y. Zhang. "DNA methylation in mammals." Cold Spring Harbor Perspectives in Biology, vol. 6, no. 5, 1 May 2014, https://doi.org/10.1101/cshperspect.a019133.

[12] Li, Hui, et al. "LSM14B is an oocyte‐specific rna‐binding protein indispensable for maternal mrna metabolism and oocyte development in mice." Advanced Science, vol. 10, no. 18, 21 Apr. 2023, https://doi.org/10.1002/advs.202300043.

[13] Li, Wei, et al. "RNA-binding proteins in regulating mRNA stability and translation: Roles and mechanisms in cancer." Seminars in Cancer Biology, vol. 86, Nov. 2023, pp. 664–677, https://doi.org/10.1016/j.semcancer.2022.03.025.

[14] Lige, Bao. "LSM FAMILY, MEMBER 14B; LSM14B." OMIM, 2024, omim.org/entry/620689#geneStructure.

[15] Liu, Tianqi, et al. "Genome-wide transcriptomics and copy number profiling identify patient-specific CNV-lncrna-mrna regulatory triplets in colorectal cancer." *Computers in*

Biology and Medicine, vol. 153, Feb. 2023, p. 106545, https://doi.org/10.1016/j.compbiomed.2023.106545.
[16] “LSM14B Protein Expression Summary - the Human Protein Atlas.” The Human Protein Atlas, www.proteinatlas.org/ENSG00000149657-LSM14B. Accessed Aug. 2025.
[17] Murphy, Robyn M., and Graham D. Lamb. “Important considerations for protein analyses using antibody based techniques: Down‐sizing western blotting up‐sizes outcomes.” The Journal of Physiology, vol. 591, no. 23, 11 Nov. 2013, pp. 5823–5831, https://doi.org/10.1113/jphysiol.2013.263251.
[18] Navarro, Miriam. “Understanding Zebrafish Development Stages | Zeclinics Cro.” ZeClinics, Apr. 2025, www.zeclinics.com/blog/understanding-zebrafish-development-stages/.
[19] Passmore, Lori A., and Jeff Coller. “Roles of mrna poly(a) tails in regulation of eukaryotic gene expression.” Nature Reviews Molecular Cell Biology, vol. 23, no. 2, 30 Sept. 2021, pp. 93–106, https://doi.org/10.1038/s41580-021-00417-y.
[20] RNA Binding Proteins – RNA Innovation, rnainnovation.med.upenn.edu/scientific-groups/rna-binding-proteins/. Accessed July 2026.
[21] Runel, Gaël, et al. “Biomechanical properties of cancer cells.” Cells, vol. 10, no. 4, 13 Apr. 2021, p. 887, https://doi.org/10.3390/cells10040887.
[22] Smyth, M S, and J H Martin. “X-ray crystallography.” Molecular Pathology, vol. 53, no. 1, Feb. 2000, pp. 8–14, https://doi.org/10.1136/mp.53.1.8.
[23] Ta, Hoang Dang, et al. “Potential therapeutic and prognostic values of LSM family genes in breast cancer.” Cancers, vol. 13, no. 19, 29 Sept. 2021, p. 4902, https://doi.org/10.3390/cancers13194902.
[24] Tie, Yan, et al. “Immunosuppressive cells in cancer: Mechanisms and potential therapeutic targets.” Journal of Hematology & Oncology, vol. 15, no. 1, 18 May 2022, https://doi.org/10.1186/s13045-022-01282-8.
[25] Wan, Yanling, et al. “LSM14B is essential for oocyte meiotic maturation by regulating maternal mrna storage and clearance.” Nucleic Acids Research, vol. 51, no. 21, 27 Oct. 2023, pp. 11652–11667, https://doi.org/10.1093/nar/gkad919.
[26] “What Is Cancer? - NCI.” National Cancer Institute, Oct. 2021, www.cancer.gov/about-cancer/understanding/what-is-cancer.
[27] Wu, Xingxing, et al. “Lsm14b controls zebrafish oocyte growth by regulating polyadenylation of the mRNA poly(A) tail.” Frontiers in Marine Science, vol. 11, 2024.
[28] Zhang, Huiru, et al. “LSM14B coordinates protein component expression in the P-body and controls oocyte maturation.” ScienceDirect, vol. 51, no. 1, Jan. 2024, pp. 48–60.
[29] Zhuang, Hongkai, et al. “Identification of LSM family members as novel unfavorable biomarkers in hepatocellular carcinoma.” Frontiers in Oncology, vol. 12, 12 May 2022, https://doi.org/10.3389/fonc.2022.871771.

When Tourism Becomes "Too much": Resident-perceived Overtourism Thresholds in Kyoto and Tokyo By Koichiro Shindo

Abstract

Overtourism is often framed as a problem of "too many tourists," yet this obscures how residents themselves judge when tourism becomes excessive. This study compares resident perceptions of overtourism in two contrasting Japanese cities: Kyoto, a compact heritage city, and Tokyo, a polycentric megacity exposed to the same national tourism boom. Using a cross-sectional survey of 101 residents (Kyoto, n = 49; Tokyo, n = 52), it examines where residents locate the overtourism threshold, whether they attribute it to tourist numbers or concentration, which perceived impacts are most closely associated with residents' sense of burden, and how residents evaluate management policies. Analyses combined t-tests, non-parametric equivalents, chi-square tests, and ordinal logistic regression. Residents of both cities defined overtourism far more as a problem of spatial and temporal concentration than of aggregate numbers. Kyoto residents were significantly more likely to feel that tourism had exceeded an acceptable level, and diverged sharply from Tokyo residents on material burdens: rising living costs, commercialization of cultural spaces, worsening transport, and unequal benefit distribution. Items measuring tourist behavior and general resident–tourist friction showed no city difference. The findings support a perception-based, context-sensitive account of the overtourism threshold, which varies across contrasting urban contexts, and point toward burden-relief policies rather than blunt caps on visitor numbers.

Introduction

Over the past decade, "overtourism" has become a key term in debates about the limits of visitor growth. It can be defined as "the impact of tourism on a destination, or parts thereof, that excessively influences perceived quality of life of citizens and/or quality of visitors' experiences in a negative way" (UNWTO, 2018, p.4). Yet the concept is inherently contested in practice. Overtourism is not simply a matter of "too many" tourists; it also depends on where and when tourism concentrates, how destinations are managed, and whose judgments define the point at which tourism becomes "too much".

In Japan, the scale of inbound tourism growth makes these questions especially urgent. The Japan National Tourism Organization reports that estimated annual international visitor arrivals reached approximately 42.7 million in 2025, setting the highest annual total on record (JNTO, 2026). This was nearly five times the 2010 level of approximately 8.6 million visitors. This surge has amplified both benefits, such as spending, jobs, and urban vibrancy, and tensions, such as crowding, rising prices, friction over etiquette, and reduced local access. However, these impacts are not evenly distributed and often concentrate in particular neighborhoods and seasons.

Despite the growth of tourism, the evidence base on how residents themselves perceive these pressures remains uneven. Within Japan, much of the resident-focused overtourism literature has centered on Kyoto and often uses qualitative, viewpoint-based, or descriptive

approaches. Examples include interview-based accounts of residents' emotional strain and distrust of tourism governance (Handler, 2026), viewpoint analyses of Kyoto's community impacts (Burtis & Wise, 2024), and municipal and media surveys documenting frustration with crowding and public transport (Kyoto City Industry and Tourism Bureau, 2025). Where multiple Japanese destinations have been compared, the comparison has typically relied on non-resident proxies such as social-media sentiment, or on descriptive multi-site surveys that record whether residents feel overtourism rather than how they define its limits. This mirrors the wider international literature, in which resident-perception studies remain dominated by single-destination case studies framed through resident social carrying capacity and social exchange theory (Muller Gonzalez et al., 2018).

As a result, there is still limited comparative evidence on how residents in structurally different Japanese urban contexts, namely a compact heritage city and a polycentric megacity, define the threshold at which tourism becomes overtourism, and on whether they attribute that threshold to total visitor numbers, spatial and temporal concentration, specific material impacts, or the perceived failure of policy. The present study addresses that gap by directly comparing resident perceptions in Kyoto and Tokyo.

This study asks where tourism crosses the threshold into overtourism by comparing perceived tipping points among residents in Kyoto and Tokyo. Kyoto is often framed as a heritage city that is vulnerable to crowding and cultural disruption, while Tokyo is a megacity assumed to absorb tourism more easily; yet both contain hotspots where time-space concentration can generate overtourism-like conditions. This study provides comparative evidence on how residents define the limits of acceptable tourism and examines how crossing those limits can make tourism growth counterproductive by disrupting residents' access, transport use, cost of living, cultural spaces, and local welfare. It also examines residents' support for policy responses and their confidence in current tourism governance, helping clarify which interventions residents support and how they evaluate the legitimacy and effectiveness of current governance.

To address these questions, the study uses a comparative survey design. The survey compares residents' perceived overtourism thresholds, causal attributions, perceived impacts, and policy attitudes across the two cities.

From tourism growth to overtourism: defining the threshold

Overtourism is best defined not as a simple “high visitor numbers” condition but as a threshold problem: destinations can accommodate tourism up to a point, after which tourism is perceived as “too much.” Although overtourism as a concept encompasses the experiences of both residents and visitors, the present study focuses specifically on residents’ perceived thresholds. UNWTO defines overtourism as tourism’s impact that “excessively influences perceived quality of life of citizens and/or quality of visitors’ experiences in a negative way” (2018, p. 4), explicitly locating the phenomenon in perceived welfare and experience rather than the number of tourists alone. In a similar spirit, Goodwin frames overtourism as a situation

where “hosts or guests, locals or visitors, feel that there are too many visitors” and where quality of life and/or the visitor experience has “deteriorated unacceptably” (2017, p. 1).

This framing matters because the core question — when does tourism become overtourism? — is contested. “Too much” depends on whose values and interests define acceptability, what is being crowded (streets, transport, housing, public services, symbolic places), and how benefits and burdens are distributed. UNWTO emphasizes that overtourism is “in essence […] the absence of good management” (2018, p. 5) and notes that congestion is often localized and shaped by capacity to manage visitors, not only by absolute volumes. In other words, thresholds are boundary-like: they mark a shift in social meaning from “welcome tourism” to “unacceptable tourism,” even if visitor numbers change only marginally.

The mechanisms behind these tipping points are therefore socio-psychological, cultural, and socio-political. To explain residents’ attitudes towards tourism development, researchers have often employed Social Exchange Theory (SET), which explains support for tourism through residents’ perceived balance of benefits and costs (Ap, 1992; Mihalič, 2020). This does not imply a purely economic calculation; rather, benefits and costs may include material, emotional, cultural, and distributive dimensions. From a social exchange perspective, residents continuously evaluate whether tourism’s benefits (income, amenities, pride) outweigh costs (stress, displacement, loss of access); when perceived costs dominate, support can flip quickly — creating a threshold dynamic. Cultural-value mechanisms also matter: irritation rises when tourism is experienced as disrespectful conduct, identity dilution, or erosion of authenticity — impacts that are not reducible to headcounts alone. Politically, thresholds become salient when grievances are organized and institutionalized (or ignored), turning tourism growth into a public conflict.

These threshold dynamics have long antecedents in classic destination models that track how resident sentiment and visitor satisfaction respond to increasing visitation. Doxey’s destination irritation index (Irrindex) conceptualizes a progression in residents’ reactions from euphoria to apathy and then towards irritation and antagonism as visitation grows and negative impacts — such as overcrowding, rising living costs, and the erosion of local atmosphere — intensify. The Irrindex is useful here because it treats “too much tourism” as a progressive deterioration in resident sentiment, rather than a single numeric capacity. This is an early articulation of resident-side threshold logic central to “overtourism.” Visitor-side thresholds are captured by the “diminishing satisfaction” logic: beyond a point, additional visitation reduces overall visitor satisfaction (Clawson & Knetsch, 1966), an idea Mihalič links directly to overtourism debates.

This motivates an empirical focus on perceived tipping points: the conditions under which residents and visitors judge tourism to have crossed from acceptable to excessive. Accordingly, this study treats overtourism as a set of perceived tipping points — acceptability limits — governing residents’ tolerance for crowding, conduct, and access, rather than as an objective visitor-volume threshold.

Two urban tourism contexts: Kyoto's heritage pressure and Tokyo's polycentric scale

Over roughly the past decade, tourism in Tokyo and Kyoto has expanded substantially, but the pace and urban consequences of that growth have differed between the two cities. In Tokyo, the number of foreign visitors increased from 11.89 million in 2015 to 24.8 million in 2024, or about 2.1 times (Tokyo Convention & Visitors Bureau, 2016; Tokyo Metropolitan Government, 2025). In Kyoto, the number of foreign tourists rose from 4.82 million in 2015 to 10.88 million in 2024, or about 2.3 times (Kyoto City, 2024; Kyoto City, 2025). Although both cities experienced strong growth, Kyoto's increase was slightly steeper, and that growth has been absorbed very differently by the two cities' urban structures. In this context, Kyoto and Tokyo represent contrasting but complementary cases for studying overtourism as a threshold phenomenon: two cities exposed to the same national tourism boom, but with different capacities to absorb it.

Kyoto is frequently characterized as a high-sensitivity heritage city, attracting visitors through iconic sites such as Kiyomizu-dera, Fushimi Inari and Kinkaku-ji, narrow historic streets within a compact urban grid, and preserved heritage districts such as Gion, Higashiyama and Arashiyama. This compactness is central to why tourism pressure in the city is felt so acutely: because the historic core occupies a small area served by a limited road and bus network, growth in visitor numbers is more likely to translate into density at a small number of sites, rather than being absorbed across the city as a whole. Kyoto City's tourism plan explicitly frames this tension, emphasizing tourism "in harmony with residents' lives" and showing that local policy is concerned with balancing visitor growth and everyday life (Kyoto City, 2021). This concern is reinforced by Kato's (2025) finding that tourism-accommodation intensity hotspots have emerged specifically in neighborhoods near the periphery of the southeastern historical center — the same area encompassing Gion and Higashiyama — rather than being distributed evenly across the city. The pressure this generates is not only spatial but material and social. On the material side, the conversion of traditional machiya townhouses into hotels and short-term rentals has been linked to gentrification and population decline in the historic core (Matsushita, 2023), while rising accommodation costs prompted the city to introduce a steeply tiered accommodation tax reaching ¥10,000 per night from March 2026 (Kyoto City Tourism Association, 2026). On the social side, city surveys report that a majority of residents feel troubled by crowded tourist areas, congested public transport, and tourists' etiquette violations (Kyoto City, 2025a) — concerns serious enough that the city has introduced limited-express sightseeing buses to separate visitor flows from the local routes residents depend on for commuting and schooling (Kyoto City Transportation Bureau, n.d.). Kyoto, therefore, provides a case in which concentrated tourism growth in a compact urban form generates cumulative, highly localized material and social pressure.

Tokyo, by contrast, can be understood as a high-capacity mega-city, and its experience of the same national tourism boom looks different as a result. The Tokyo Tourism Data Catalog notes that visitors, both Japanese and international, spread across multiple wards — Shinjuku, Chiyoda, Shibuya, Minato, Chuo and Taito among them — rather than concentrating in a single

historic core (Tokyo Metropolitan Government, 2025). This suggests that tourism pressure in Tokyo is substantial in aggregate but structurally dispersed: the city's larger transport network and broader urban scale allow it to absorb comparable visitor growth with less citywide strain than Kyoto experiences. Where friction does arise, it tends to be sharply localized in time and place rather than pervasive — most visibly in Shibuya, where recurring crowding and disorder around Halloween and New Year have prompted the ward to introduce and then extend ordinances restricting public drinking (Shibuya Ward, 2024). This pattern suggests that in a polycentric city, "too much tourism" is less a condition of everyday life than an episodic event tied to specific hotspots and calendar moments.

Read together, the two cities' recent trajectories point to the central puzzle this study investigates. Kyoto and Tokyo have experienced comparable — even proportionally larger, in Kyoto's case — growth in visitor numbers, yet the character of the resulting pressure diverges sharply: appearing more diffuse and event-specific in Tokyo, concentrated and materially pervasive in Kyoto.

Managing overtourism: policy tools, soft measures and implementation limits

Policy responses to overtourism in Japan span a familiar international toolkit — transport management, fiscal measures such as accommodation taxes, access management, and behavioral codes — yet their effectiveness often falls short of expectations. Japan's response combines national policy with local measures in Kyoto and Tokyo. Overall, however, the policy mix has relied more on soft management than on strict control. The 2024 White Paper on Tourism in Japan emphasizes sustainable tourism through support for local governments and Destination Management Organizations (DMOs), improvements in visitor reception, multilingual guidance, anti-litter measures, and efforts to disperse tourism demand. It also reports that around 70% of international guest nights are concentrated in Japan's three major metropolitan areas, indicating that inbound tourism pressure is geographically uneven at the national scale (Japan Tourism Agency, 2024). This uneven distribution is especially pronounced in Tokyo and Kyoto, the two cities examined in this study. Japan has therefore framed overtourism as a sustainability and governance challenge, necessitating specific policy interventions.

Kyoto provides a case of how cities have moved beyond generic marketing and communication responses toward more interventionist measures. Kyoto uses transport management systems to reduce pressure in specific tourist hotspots and during specific times, providing congestion forecasts, live camera information, and route guidance through tools such as the Kyoto Travel Congestion Forecast and Kyoto Smart Navi, so that visitors can avoid hotspots during peak times. Kyoto has also sought to ease crowding on public transport by increasing bus frequency on major sightseeing routes and introducing limited express buses. In terms of behavioral codes, Kyoto promotes responsible travel by encouraging visitors to respect local communities, scenic landscapes, and cultural heritage. On the fiscal side, Kyoto revised its accommodation tax effective from March 1, 2026, adopting a steeper tiered structure ranging from ¥200 to ¥10,000 per person per night depending on the room rate (Kyoto City Tourism

Association, 2026). This explicitly positions revenue as a resource for managing impacts and sustaining the city's tourism–community balance.

Tokyo's response has been more centered on behavioral guidance and smoother visitor circulation than on direct restrictions. Through GO TOKYO's guidance for responsible travel, visitors are encouraged to learn local rules, respect communities, preserve the environment, and behave appropriately in places such as shrines, temples, public baths, and public transportation. Tokyo also uses manners campaigns, tourist information, and guide services to reduce friction in crowded urban areas. Although Tokyo also has an accommodation tax, its recent overtourism response has been more focused on information provision and behavioral management than on place-based restrictions (Tokyo Convention & Visitors Bureau, n.d.).

However, policies may underperform in practice for a number of reasons. The core problem is that tourist demand is concentrated in a small number of famous districts, attractions, and peak seasons, yet many measures still rely on information provision, moral suasion, and voluntary dispersion rather than enforceable limits. Congestion forecasts, etiquette campaigns, and route guidance may reduce friction and improve visitor behavior at the margin, but they do not necessarily reduce the number of people gathering in the same hotspot at the same time. Accommodation taxes are also limited as a corrective tool because they generate revenue more reliably than they change tourist behavior unless charges are high or strongly targeted. As a result, the policies may look comprehensive on paper, but they often remain too soft to alter the underlying geography of tourism concentration in a decisive way (Japan Tourism Agency, 2024; Nepal & Nepal, 2019).

The above observations raise the four questions this study addresses. First, do residents of the two cities in fact perceive different thresholds for when tourism becomes excessive, given their differing exposure to concentrated pressure (RQ1)? Second, is that pressure better understood by residents as a problem of sheer visitor numbers or of where and when those visitors concentrate (RQ2)? Third, which specific impacts — crowding, cost of living, cultural commercialization, or the distribution of tourism's benefits — appear more central to residents' sense that tourism has become a burden, and do these differ between the two cities (RQ3)? Finally, do residents' policy preferences and their trust in current tourism governance track these differences in material experience, or does skepticism toward management extend even to the less-burdened city (RQ4)? The remainder of this study addresses these questions through a comparative resident survey.

Methodology

Research design

This study employed a cross-sectional, comparative survey design to examine residents' perceptions of tourism and overtourism in two Japanese cities, Tokyo and Kyoto. A quantitative approach was adopted to address the study's four research questions, which concern (a) the perceived threshold of overtourism and whether it differs between the two cities, (b) whether

overtourism is understood primarily as a problem of tourist numbers or of tourist concentration, (c) residents' perceptions of the impacts, benefits, and costs of tourism, and (d) residents' support for and evaluation of tourism management policies. Kyoto and Tokyo were selected as contrasting cases for comparison; the detailed rationale for this choice is developed in the Introduction.

Sampling and recruitment

Participants were residents of Tokyo and Kyoto, recruited through a self-administered online questionnaire. Recruitment was carried out by distributing printed leaflets displaying a QR code linking to the questionnaire, which was hosted on Google Forms. The lead researcher personally distributed 7,000 leaflets by hand. To capture variation in residents' proximity to tourist activity and exposure to tourism, leaflets were distributed in roughly equal proportions across three zones in each city: the center of the major tourist areas, their immediate surroundings, and locations well removed from them. This strategy was intended to ensure that the sample was not confined to residents living within tourist hotspots, thereby producing meaningful variation on the proximity and frequency-of-exposure measures.

Recruitment thus followed a convenience sampling approach, since participation depended on residents encountering a flyer and choosing to respond. A total of 101 valid responses were obtained (Kyoto, n = 49; Tokyo, n = 52), corresponding to a participation yield of 1.4% relative to the number of leaflets distributed.

Data collection

Data was collected over a one-month period, from 1 to 30 May 2026, using Google Forms. Responses were recorded anonymously, and the questionnaire could be completed on any internet-connected device.

Survey instrument and measures

The questionnaire was administered, and all responses were collected and analyzed, in the original Japanese; the English item wording reported in this paper is provided solely for reporting purposes and reflects the lead researcher's own translation of the original Japanese items. Back-translation was not conducted, which is an acknowledged limitation.

The first section collected background information: age group (under 18; 18–29; 30–39; 40–49; 50–59; 60–69; 70 or older), gender (male, female, prefer not to answer), city of residence (Tokyo or Kyoto), length of residence in the current city (under 1 year; 1 to under 3; 3 to under 5; 5 to under 10; 10 to under 20; 20 years or more), main mode of daily transport (walking, bicycle, motorcycle, private car, bus, train, taxi, other), approximate distance from home to a representative tourist area (under 1 km; 1–3; 3–5; 5–10; 10–30; 30 km or more), and frequency of visiting tourist-heavy areas (three or more times per week; one to two times per week; one to three times per month; one to three times per year; rarely; never). The representative tourist areas were specified as the Shibuya Scramble Crossing area for Tokyo and the Gion–Kiyomizu area

for Kyoto. These areas were used as reference points for measuring residents' approximate proximity to tourist activity, rather than as exhaustive representations of tourism activity in each city; both cities contain multiple additional areas of concentrated tourist activity not captured by this single reference point.

The remaining four sections measured the study's key constructs using a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). These sections covered the perceived overtourism threshold and the numbers-versus-concentration distinction; perceived impacts (crowding and loss of access, norm violations, and cultural impacts); perceived benefits and costs; and policy support and policy evaluation. The constructs and their corresponding items are summarized in Table 1, grouped by the research question (RQ1–RQ4) they address.

The items were developed by the researcher on the basis of the overtourism and resident-attitude literature. Because the instrument was newly constructed for this study, no formal pilot validation was conducted.

All Likert-scale items were treated as ordinal-level measurements. This reflects the nature of Likert responses, which represent ordered categories without assumed equal intervals, and it provides the basis for using non-parametric tests alongside parametric procedures in the analysis (see Data analysis).

Table 1. Constructs and survey items

Construct / section	Item	Item wording (researcher's translation)
RQ1 — Overtourism threshold		
Threshold	RQ1	Tourism in the city where I live currently exceeds an acceptable level.
Threshold (QoL)	RQ1B	Tourism becomes overtourism when it significantly lowers residents' quality of life.
RQ2 — Tourist numbers versus concentration		
Cause: numbers	RQ2A	The main problem of overtourism is the large number of tourists itself.
Cause: concentration	RQ2B	The main problem of overtourism is not the total number of tourists itself, but their concentration in specific places or at specific times.
RQ3 — Perceived impacts, benefits, and costs		
Crowding / access	RQ3A	Due to the increase in tourists, daily travel has become less convenient.
Norm violations	RQ3B	I often observe inappropriate behavior by tourists.
Norm violations	RQ3C	Friction between residents and tourists is common in tourist-heavy areas.
Cultural impacts	RQ3D	The area's original atmosphere and character are changing to suit tourism.
Cultural impacts	RQ3E	Local cultural spaces feel commercialized for tourists rather than serving residents.
Benefits	RQ3F	Tourism creates employment and income opportunities.

Benefits	RQ3G	Tourism increases the recognition and attractiveness of the area.
Benefits	RQ3H	Overall, tourism benefits my local community.
Costs	RQ3I	Tourism contributes to rising prices and living costs.
Costs / distribution	RQ3J	The benefits of tourism are not distributed equally among residents.
RQ4 — Policy support and evaluation		
Policy support	RQ4A	I support policies that limit visitor numbers in highly congested areas.
Policy support	RQ4B	I support policies that redistribute tourists to less crowded areas.
Policy evaluation	RQ4C	Pricing measures such as accommodation taxes or entry fees are effective.
Policy support	RQ4D	Improvements in public transportation and infrastructure can reduce overtourism.
Policy evaluation	RQ4E	Current government tourism policies are functioning effectively.
Policy evaluation	RQ4F	I trust local authorities to manage tourism impacts effectively.

Note. Items are grouped by research question, following the numbering used in the Results.

Data analysis

All statistical analyses were conducted in jamovi. The analysis proceeded in three stages. First, descriptive statistics were used to summarize the sample and response distributions; frequencies and percentages were reported for categorical background variables (gender, age group, length of residence, main mode of daily transport, distance from home to a representative tourist area, and frequency of visiting tourist-heavy areas). Second, inferential analyses were conducted to address the four research questions: independent samples t-tests compared mean agreement between Kyoto and Tokyo residents on each item, and a paired samples t-test compared the relative weight residents placed on tourist numbers versus tourist concentration (RQ2). Third, supplementary analyses were used where appropriate to test the robustness of these comparisons and to model place of residence as a predictor — namely non-parametric equivalents (Mann-Whitney U, Wilcoxon signed-rank, and Kruskal-Wallis H tests), chi-square tests of independence on the response distributions, and ordinal logistic regression with place of residence as the predictor of residents' ordinal responses. Differences between Kyoto and Tokyo residents were tested using independent-samples t-tests, with the Mann-Whitney U test reported alongside as a non-parametric check, given the ordinal nature of the items and observed violations of the assumptions of normality and homogeneity of variance. Associations between city of residence and categorical response patterns were examined using chi-square tests of independence based on contingency tables. Differences across groups were also tested with the Kruskal–Wallis H test. Supplementary tests, such as an ordinal logistic regression, were used to model each ordinal outcome as a function of city of residence, with odds ratios reported where

relevant to aid interpretation; the direction of each coefficient was interpreted according to the city coding and reference category used in jamovi. Within-sample comparisons of the two numbers-versus-concentration items were tested using a paired-samples t-test, with the Wilcoxon signed-rank test reported as a non-parametric check. An alpha level of .05 was applied throughout.

Ethical considerations

Participation in the study was voluntary and anonymous, and no personally identifiable information was collected. Before beginning the questionnaire, participants were presented with an information statement describing the purpose and academic nature of the study and identifying the lead researcher (Koichiro Shindo). The statement explained that responses were collected anonymously, that participants could withdraw at any point before final submission without any consequence, and that continuing with the questionnaire would be taken as consent to participate (implied consent). A contact email address was provided for any questions.

The collected data were stored on a password-protected device accessible only to the researcher, contained no personally identifiable information, and were scheduled to be deleted six months after completion of the study.

Results

Descriptive analysis

A total of 101 residents completed the survey, comprising 49 respondents from Kyoto and 52 from Tokyo. The sample was balanced by gender, with 50 respondents identifying as female, 50 as male, and 1 preferring not to answer.

In terms of age, respondents were distributed across all age bands, with the 30–39 (n = 23) and 50–59 (n = 23) groups being the largest, followed closely by 18–29 (n = 20) and 40–49 (n = 14). Smaller numbers fell into the 60–69 (n = 11), 70-or-older (n = 6), and under-18 (n = 4) groups. The age profile differed somewhat by city: in Kyoto the largest group was 30–39 (n = 15), whereas in Tokyo respondents clustered in the 18–29 (n = 14) and 50–59 (n = 14) groups.

The sample consisted predominantly of long-term residents. A majority had lived in their current city for 20 years or more (n = 43), and a further 33 had lived there for 10 to less than 20 years; together these two groups accounted for roughly three-quarters of the sample. Shorter-term residents were comparatively rare (5 to less than 10 years, n = 16; 3 to less than 5 years, n = 6; 1 to less than 3 years, n = 3).

The train was by far the most common mode of daily transport (n = 66, 65%), followed by bus (n = 14), bicycle (n = 7), walking (n = 6), and car (n = 5). Reliance on the train was especially pronounced in Tokyo, where 42 of 52 residents reported it as their main mode, compared with 24 of 49 in Kyoto; bus use was correspondingly more common in Kyoto (n = 13) than in Tokyo (n = 1).

With respect to proximity to representative tourist areas (the Shibuya Scramble Crossing area for Tokyo and the Gion–Kiyomizu area for Kyoto), respondents were spread across the full range of distances. The most common distances were 3–5 km (n = 27) and 5–10 km (n = 27), followed by 1–3 km (n = 20) and 10–30 km (n = 17), with smaller numbers living less than 1 km (n = 8) or 30 km or more (n = 2) away. Kyoto residents tended to live closer to the tourist core (3–5 km was the most common band, n = 17, and 5 lived within 1 km), whereas Tokyo residents were more dispersed toward the middle and outer distances (5–10 km, n = 17; 10–30 km, n = 12). This spread reflects the recruitment strategy, in which flyers were distributed across the center, surroundings, and more distant areas of each city.

Finally, frequency of visiting tourist-heavy areas differed markedly between the two cities. Overall, the most common responses were visiting one to three times a year (n = 29) and rarely (n = 25). Tokyo residents reported visiting far more frequently than Kyoto residents: 12 of 52 Tokyo residents visited three or more times a week, compared with only 3 of 49 in Kyoto, and no Tokyo resident reported never visiting, whereas 6 Kyoto residents did. Kyoto residents were concentrated instead in the infrequent categories (one to three times a year, n = 18; rarely, n = 17).

Inferential analyses

RQ1: Where is the overtourism threshold in each city? Do Kyoto and Tokyo differ in the overtourism threshold?

> *RQ1: Tourism in the city where I live currently exceeds an acceptable level.*
> *RQ1B: I think tourism becomes overtourism when it significantly lowers residents' quality of life.*

The first research question examined whether residents in Kyoto and Tokyo differed in their perception that tourism in their city has already exceeded an acceptable level. An independent sample t-test showed a statistically significant difference between the two cities, $t(99) = 2.85$, $p = .005$, with a moderate-to-strong effect size, Cohen's $d = 0.57$. A Mann-Whitney U test also supported this difference, $U = 896$, $p = .007$. These results indicate that Kyoto residents were significantly more likely than Tokyo residents to agree that tourism in their city already exceeds an acceptable level.

This pattern is also visible in the response distribution. In Kyoto, 13 out of 49 respondents strongly agreed that tourism had exceeded an acceptable level, compared with only 5 out of 52 respondents in Tokyo. A chi-square test further showed that the distribution of responses differed significantly by city, $\chi^2(4, N = 101) = 10.8$, $p = .029$. Overall, these findings suggest that the perceived overtourism threshold has been crossed more clearly in Kyoto than in Tokyo.

Table 2. Distribution of responses to "Tourism in the city where I live currently exceeds an acceptable level" by city

Place of residence	Strongly agree	Somewhat agree	Neither	Somewhat disagree	Strongly disagree	Total
Kyoto	13	21	5	10	0	49
Tokyo	5	22	3	17	5	52
Total	18	43	8	27	5	101

Note. $\chi^2(4, N = 101) = 10.8, p = .029$

The second part of RQ1 examined whether residents differed in their view that overtourism is reached when residents' quality of life is significantly reduced. In this case, there was no statistically significant difference between Kyoto and Tokyo residents, t(99) = -0.19, p = .850, Cohen's d = -0.04. The Mann-Whitney U test also showed no significant difference, U = 1271, p = .985. This suggests that residents in both cities similarly recognize quality-of-life decline as an important threshold of overtourism.

RQ2: Is overtourism seen as a problem of tourist numbers or tourist concentration?

RQ2: The main problem of overtourism is the large number of tourists itself.

RQ2B: The main problem of overtourism is not the total number of tourists itself, but their concentration in specific places or at specific times.

RQ2 examined whether overtourism is understood mainly as a problem of overall tourist numbers or as a problem of tourist concentration in specific areas. For the item measuring overall tourist numbers, Kyoto residents reported significantly stronger agreement than Tokyo residents, t(99) = 2.41, p = .018, Cohen's d = 0.48. The Mann-Whitney U test also showed a significant difference, U = 929, p = .012. This indicates that Kyoto residents are more likely than Tokyo residents to view overtourism as a problem caused by the overall number of tourists.

However, for the item measuring tourist concentration, there was no statistically significant difference between Kyoto and Tokyo residents, t(99) = -0.63, p = .533, Cohen's d = -0.13. The Mann-Whitney U test also showed no significant difference, U = 1214, p = .648. This suggests that residents in both cities similarly recognize tourist concentration as a key feature of overtourism.

To compare the relative importance of numbers and concentration within the full sample, a paired samples t-test was conducted. The mean score for tourist concentration was much higher (M = 4.40, SD = 0.76) than the mean score for overall tourist numbers (M = 2.68, SD = 1.06). This difference was statistically significant, t(100) = -14.7, p < .001. The Wilcoxon test also supported this result, W = 60.0, p < .001. These findings show that residents are much more likely to define overtourism as a problem of tourist concentration than as a problem of tourist numbers alone.

Taken together, the results suggest that Kyoto residents perceive overtourism as having crossed an acceptable threshold more strongly than Tokyo residents. However, when defining

what causes overtourism, residents in both cities place greater emphasis on the concentration of tourists in specific areas than on the total number of tourists overall. Kyoto residents are more likely than Tokyo residents to emphasize overall tourist numbers, but there is no significant city-level difference in views about concentration. This indicates that overtourism is understood less as a simple issue of "too many tourists" and more as a spatial problem involving where tourists gather and how that concentration affects residents' daily lives.

RQ3: How are the impacts, benefits, and costs of tourism perceived, and do they differ between cities?

Having established where the overtourism threshold lies (RQ1) and how residents understand its underlying cause (RQ2), RQ3 turns to how residents actually experience the consequences of tourism in their daily lives. Its overall aim is to map residents' perceptions of the concrete impacts, benefits, and costs of tourism, and to test whether these perceptions differ between Kyoto and Tokyo. The items are organized into five clusters: crowding and loss of access (RQ3A), norm violations (RQ3B and RQ3C), cultural impacts (RQ3D and RQ3E), the perceived benefits of tourism (RQ3F–RQ3H), and its perceived costs and their distribution (RQ3I and RQ3J). Each item was analyzed with the same battery of tests (independent samples t-test, Mann-Whitney U, chi-square, Kruskal-Wallis, and ordinal logistic regression), and the section concludes by drawing out the overall pattern that distinguishes the two cities. *Note.* For the RQ3 items, responses were entered so that the ordinal logistic regressions model increasing agreement; a positive "Tokyo – Kyoto" coefficient therefore indicates stronger agreement among Kyoto than Tokyo residents. The RQ4 policy models reported later were entered with the reverse category order, so their coefficient signs should be read against that coding rather than by the same rule.

Crowding and loss of access (RQ3A: "Due to the increase in tourists, transportation has become worse.")

For RQ3A, residents were asked whether transportation had worsened due to the increase in tourists. Kyoto residents reported significantly stronger agreement than Tokyo residents. An independent sample t-test revealed a significant difference with a large effect size, $t(99) = 4.57$, $p < .001$, Cohen's $d = 0.91$, and the Mann-Whitney U test supported this result, $U = 634$, $p < .001$. The distribution of responses also differed significantly by city, $\chi^2(4, N = 101) = 25.2$, $p < .001$: in Kyoto, 30 of 49 residents strongly agreed that transportation had worsened, compared with only 7 of 52 residents in Tokyo. A Kruskal-Wallis test produced the same conclusion, $\chi^2(1) = 20.5$, $p < .001$, $\varepsilon^2 = 0.205$. Finally, an ordinal logistic regression showed that city significantly predicted responses, with Kyoto residents far more likely than Tokyo residents to agree that transportation had worsened ($b = 1.80$, $SE = 0.40$, $Z = 4.50$, $p < .001$; $OR \approx 6.05$). Taken together, these results indicate that perceived transport-related strain from tourism is substantially greater in Kyoto than in Tokyo. This fits a daily-life friction interpretation of overtourism, in which the phenomenon is experienced primarily through everyday mobility

constraints — a tangible, high-frequency impact — rather than only through more symbolic or abstract concerns.

Norm violations (RQ3B: inappropriate behavior by tourists; RQ3C: friction between residents and tourists)

For RQ3B (observing inappropriate behavior by tourists), there was no significant difference between Kyoto and Tokyo residents across all tests, $t(99) = 0.99$, $p = .327$; $U = 1148$, $p = .355$; $\chi^2(4) = 7.44$, $p = .114$; Kruskal-Wallis $\chi^2(1) = 0.86$, $p = .353$; ordinal regression $b = 0.35$, $p = .352$. The perception of tourists' inappropriate behavior is shared similarly across both cities.

For RQ3C (friction between residents and tourists in crowded areas), there was no significant difference between Kyoto and Tokyo residents across all tests, $t(99) = 0.47$, $p = .637$; $U = 1211$, $p = .650$; $\chi^2(4) = 1.24$, $p = .872$; Kruskal-Wallis $\chi^2(1) = 0.21$, $p = .647$; ordinal regression $b = 0.17$, $p = .646$. Perceived friction in tourist-dense areas is similar in both cities.

For both norm-violation items — observing inappropriate behavior (RQ3B) and friction between residents and tourists (RQ3C) — there were no significant differences between Kyoto and Tokyo residents on any test (all $p > .11$). Perceptions of tourist-related norm violations appear to be shared across the two cities, in contrast to the infrastructure strain seen in RQ3A.

Cultural impacts (RQ3D: loss of original local character; RQ3E: commercialization of cultural spaces)

For RQ3D (feeling that the area's original atmosphere and character have been lost), there was no significant difference between Kyoto and Tokyo residents on any test, $t(99) = -0.32$, $p = .747$; $U = 1241$, $p = .810$; $\chi^2(4) = 0.80$, $p = .939$; Kruskal-Wallis $\chi^2(1) = 0.06$, $p = .808$; ordinal regression $b = -0.09$, $p = .807$. Both cities report a similar sense of lost local character. One possibility is that this item is relatively broad, so that respondents may have had different reference points in mind; it may therefore fail to discriminate between the cities even where more specific cultural pressures do differ, as the next item (RQ3E) suggests.

For RQ3E, residents were asked whether local cultural spaces are being commercialized. Kyoto residents reported markedly stronger agreement than Tokyo residents, with the largest effect observed for any RQ3 item. The independent samples t-test was significant with a very large effect size, $t(99) = 7.48$, $p < .001$, Cohen's $d = 1.49$, and the Mann-Whitney U test was consistent with this pattern, $U = 431$, $p < .001$. The response distribution differed sharply by city, $\chi^2(4, N = 101) = 38.5$, $p < .001$: in Kyoto, 34 of 49 residents strongly agreed, compared with only 9 of 52 in Tokyo, where the majority disagreed. The Kruskal-Wallis test produced the same conclusion, $\chi^2(1) = 36.2$, $p < .001$, and an ordinal logistic regression showed that city strongly predicted responses, with Kyoto residents far more likely than Tokyo residents to perceive commercialization ($b = 2.63$, $SE = 0.45$, $Z = 5.87$, $p < .001$; $OR \approx 13.9$). Perceived commercialization of cultural spaces is thus far more pronounced in Kyoto than in Tokyo. This is one of the study's strongest findings and is consistent with a heightened heritage-city sensitivity

in Kyoto; it may also help to explain why Kyoto residents more often judged tourism to be beyond an acceptable level in RQ1, a link developed further in the Discussion.

Benefits (RQ3F: employment and income; RQ3G: recognition and attractiveness; RQ3H: overall community benefit)

For RQ3F (tourism creating employment and income opportunities), there was no significant difference between Kyoto and Tokyo residents on any test, $t(99) = 1.59$, $p = .115$; $U = 1098$, $p = .169$; $\chi^2(2) = 5.18$, $p = .075$; Kruskal-Wallis $\chi^2(1) = 1.90$, $p = .168$; ordinal regression $b = 0.56$, $p = .167$. Residents in both cities similarly recognize the economic benefits of tourism.

For RQ3G, residents were asked whether tourism increases the recognition and attractiveness of their city. Kyoto residents reported significantly stronger agreement than Tokyo residents, $t(99) = 2.71$, $p = .008$, Cohen's $d = 0.54$, with the Mann-Whitney U test supporting this, $U = 969$, $p = .015$. The distribution differed by city at the threshold of significance, $\chi^2(3, N = 101) = 7.83$, $p = .050$, with more Kyoto residents strongly agreeing (14 of 49 vs. 7 of 52) and more Tokyo residents disagreeing. A Kruskal-Wallis test produced the same conclusion, $\chi^2(1) = 5.96$, $p = .015$, and an ordinal logistic regression likewise showed a significant effect of city ($b = 1.04$, $SE = 0.43$, $Z = 2.42$, $p = .016$). While residents in both cities generally acknowledge this benefit ("somewhat agree" was most common in both), the positive perception is somewhat stronger in Kyoto. Kyoto residents thus perceive both a higher upside (greater recognition and attractiveness) and, as shown below, substantially higher costs — a pattern consistent with a social-exchange profile in which strong perceived benefits coexist with strong perceived burdens.

For RQ3H (overall, tourism benefits the local community), there was no significant difference between Kyoto and Tokyo residents on any test, $t(99) = 0.85$, $p = .399$; $U = 1156$, $p = .393$; $\chi^2(4) = 2.58$, $p = .630$; Kruskal-Wallis $\chi^2(1) = 0.74$, $p = .391$; ordinal regression $b = 0.32$, $p = .388$. Residents in both cities hold similar overall views on whether tourism benefits their community. Notably, then, despite the marked city differences on specific benefit and cost components reported elsewhere in this section, the net judgment of whether tourism benefits one's community does not differ substantially between Kyoto and Tokyo.

Across the three benefit-related items, residents in both cities generally acknowledged the benefits of tourism. There were no significant city differences for employment and income opportunities (RQ3F, $p = .115$) or for overall community benefit (RQ3H, $p = .399$), with residents in both cities viewing these similarly. By contrast, Kyoto residents more often agreed that tourism increases the city's recognition and attractiveness (RQ3G, $p = .008$, $d = 0.54$). Thus, while the existence of tourism's benefits is broadly accepted, Kyoto residents are somewhat more attuned to its positive visibility effects.

Costs (RQ3I: rising prices and living costs)

For RQ3I, residents were asked whether tourism contributes to rising prices and living costs. This item produced the largest city difference of any RQ3 measure. Kyoto residents agreed

far more strongly than Tokyo residents, $t(99) = 9.90$, $p < .001$, Cohen's $d = 1.97$ — an exceptionally large effect — and the Mann-Whitney U test was consistent with this pattern, $U = 224$, $p < .001$. The distributions were almost mirror images: in Kyoto, 35 of 49 residents strongly agreed (with a further 9 somewhat agreeing), whereas in Tokyo 31 of 52 residents disagreed and only one strongly agreed, $\chi^2(4, N = 101) = 58.4$, $p < .001$. The Kruskal-Wallis test produced the same conclusion, $\chi^2(1) = 54.9$, $p < .001$, $\varepsilon^2 = 0.549$, and an ordinal logistic regression showed that city strongly predicted responses, with Kyoto residents far more likely than Tokyo residents to agree ($b = 3.66$, $SE = 0.54$, $Z = 6.77$, $p < .001$; OR ≈ 39). Perceived tourism-related inflation in prices and living costs is thus dramatically higher in Kyoto than in Tokyo. Because affordability is such an immediate and powerful concern, this perceived cost-of-living pressure is likely to be one of the key mechanisms behind the stricter "acceptable level" assessments observed among Kyoto residents in RQ1, a point taken up in the Discussion.

Unequal distribution of benefits (RQ3J: benefits are not shared equally among residents)

For RQ3J (the benefits of tourism are not distributed equally among residents), Kyoto residents agreed more strongly than Tokyo residents, $t(99) = 2.49$, $p = .015$, Cohen's $d = 0.50$; $U = 976$, $p = .022$; Kruskal-Wallis $\chi^2(1) = 5.23$, $p = .022$; ordinal regression $b = 0.92$, $p = .023$. The chi-square test was non-significant ($\chi^2(3) = 5.97$, $p = .113$) because residents in both cities overwhelmingly agreed (strongly or somewhat); the difference lies in the intensity of agreement, with more Kyoto residents strongly agreeing (31 of 49 vs. 22 of 52). Notably, this is a near-universal concern, but it is felt most acutely in Kyoto.

Taken together, the RQ3 items reveal a clear and consistent pattern in how overtourism's impacts are perceived across the two cities. The differences between Kyoto and Tokyo were not uniform but concentrated in a specific type of impact. Kyoto residents reported significantly stronger agreement on every item involving a direct, material burden on daily life: rising prices and living costs (RQ3I, $d = 1.97$), the commercialization of local cultural spaces (RQ3E, $d = 1.49$), worsening transportation (RQ3A, $d = 0.91$), and the unequal distribution of tourism's benefits (RQ3J, $d = 0.50$). They were also somewhat more likely to recognize tourism's positive effect on the city's recognition and attractiveness (RQ3G, $d = 0.54$). In contrast, the two cities did not differ on the more attitudinal or behavioral items: observing inappropriate behavior by tourists (RQ3B), friction between residents and tourists (RQ3C), the loss of the area's original atmosphere (RQ3D), and the broad economic benefits of tourism in terms of employment, income, and overall community benefit (RQ3F, RQ3H). In other words, residents of both cities hold similar attitudes about tourists and broadly agree that tourism brings benefits, but they diverge sharply on the extent to which tourism imposes concrete costs on everyday life — and these costs are felt far more acutely in Kyoto. This pattern is consistent with the findings for RQ1, where Kyoto residents were more likely to feel that tourism had already exceeded an acceptable threshold: the perception that "there are too many tourists" appears to be associated less with tourists' conduct or abstract cultural change than with the tangible economic and infrastructural pressures that tourism places on residents' daily lives.

RQ4: Do residents differ in their attitudes toward tourism management policies?

RQ4 examined residents' attitudes toward various tourism management policies. The clearest city difference emerged for support of public transportation and infrastructure improvements (RQ4D), which Kyoto residents endorsed significantly more strongly than Tokyo residents, $t(99) = 2.69$, $p = .008$, Cohen's $d = 0.54$; $U = 826$, $p = .001$; $\chi^2(4, N = 101) = 16.8$, $p = .002$; ordinal regression $b = -1.30$, $p = .001$. This mirrors the transport-related strain reported more acutely in Kyoto (RQ3A). Support for dispersing tourists to less crowded areas (RQ4B) showed a weak difference in the same direction, but the parametric and non-parametric tests diverged ($t(99) = 2.10$, $p = .038$; $U = 1049$, $p = .075$) and the result should be interpreted with caution. For all remaining items, there were no significant differences between the two cities: support for limiting tourist numbers (RQ4A, $p = .232$), price-based measures such as accommodation taxes (RQ4C, $p = .545$), the belief that current government tourism policies are functioning well (RQ4E, $M_{Tokyo} = 2.56$, $SD = 0.96$; $M_{Kyoto} = 2.31$, $SD = 0.90$; $p = .176$), and trust in local administrative authorities (RQ4F, $M_{Tokyo} = 2.83$, $SD = 1.10$; $M_{Kyoto} = 2.59$, $SD = 1.15$; $p = .296$). Overall, although Kyoto residents experience tourism's material costs more acutely, this is reflected mainly in a stronger desire for infrastructure improvement rather than in greater support for restrictive or price-based measures. Moreover, the absence of significant city differences in perceived policy effectiveness (RQ4E) and trust in authorities (RQ4F) suggests that skepticism toward current tourism governance is not unique to Kyoto but is shared across both urban contexts.

Discussion

This study set out to compare how residents of Kyoto and Tokyo perceive overtourism, and the findings reveal a consistent and theoretically meaningful pattern. Across the four research questions, the two cities did not differ uniformly. Instead, residents of both cities did not differ significantly on several attitudinal items and broadly agreed that tourism brings benefits, yet they diverged sharply on the extent to which tourism imposes tangible, material costs on everyday life — costs that were felt far more acutely in Kyoto. Kyoto residents were more likely to feel that tourism had already crossed an acceptable threshold (RQ1), reported substantially stronger agreement on costs, cultural impacts, and crowding (RQ3), and expressed stronger support for transport and infrastructure improvements (RQ4). At the same time, residents of both cities defined overtourism far more as a problem of spatial and temporal concentration than of aggregate tourist numbers (RQ2). Taken together, these results suggest that the perceived overtourism threshold is reached not through abstract cultural change or tourist conduct, but through the concrete economic and infrastructural pressures that concentrated tourism places on residents' daily lives, and that these pressures are shaped by the very different urban forms of the two cities.

RQ1: The overtourism threshold as a perception-based judgment

The finding that Kyoto residents were significantly more likely than Tokyo residents to feel that tourism has exceeded an acceptable level, despite no city difference in the abstract definition of overtourism as a decline in quality of life, supports a perception-based rather than indicator-based understanding of the overtourism threshold. The threshold is not a fixed visitor count but a subjective judgment that varies with local context. This is consistent with Koens et al. (2018), who argue that overtourism is multidimensional, that its impacts are localized rather than city-wide, and that it is not synonymous with mass tourism, and with the UNWTO (2018) emphasis on resident perceptions as central to defining and managing urban tourism growth. Doxey's (1975) Irrindex offers a useful interpretive lens here: as tourism saturation rises, host attitudes are theorized to move through euphoria, apathy, annoyance and antagonism. The pattern is consistent with the kind of irritation dynamic described by Doxey (1975), although the present data cannot locate Kyoto definitively within a particular stage of the Irrindex. Kyoto's manner campaigns, photography restrictions, and the local discourse of "tourism pollution" (kankō kōgai) are suggestive of this dynamic, but they are not direct evidence of residents' position on the index. The Irrindex should, however, be treated as a heuristic rather than a predictive law, given its well-documented assumptions of community homogeneity and unidirectional change and its limited empirical grounding (Sharpley, 2014).

RQ2: Concentration, not numbers, and the challenge to the "too-many-tourists" narrative

One of the clearest results of the study is that residents of both cities defined overtourism overwhelmingly as a problem of tourist concentration in specific places and at specific times, rather than of the total number of tourists ($M = 4.40$ versus $M = 2.68$; paired $t(100) = -14.7$, $p < .001$). This directly challenges the popular "too-many-tourists" narrative and reframes overtourism as a spatial and temporal phenomenon. This does not mean that overall visitor numbers are irrelevant, but rather that their impact is mediated by where, when, and how those visitors are distributed across the city. The result is best understood through the concept of social or perceptual carrying capacity, which concerns the point at which residents begin to feel crowded or displaced, as distinct from a destination's physical capacity (Muler Gonzalez, Coromina, & Galí, 2018; O'Reilly, 1986). Overtourism, in this framing, is produced by honeypot effects — the funnelling of visitors into a small number of iconic sites and streets — rather than by aggregate visitor volumes spread evenly across a city (Koens, Postma and Papp, 2018). This interpretation has direct policy relevance. If the problem is concentration rather than numbers, then the most appropriate responses are those that manage visitor flows across space and time — dispersal to less-visited areas, off-peak and early-morning promotion, real-time congestion forecasting and capacity management at hotspots — rather than blunt caps on total arrivals. That Kyoto residents nonetheless emphasized overall numbers somewhat more than Tokyo residents is itself consistent with the urban-form argument developed below: in a compact city, high concentration and high aggregate numbers are experienced as the same daily reality.

RQ3: Material burdens, urban form, and the divergence between the two cities

The RQ3 findings form the empirical core of the study. Kyoto residents reported significantly stronger agreement across a cluster of material-burden items — rising prices and living costs (d = 1.97), commercialization of cultural spaces (d = 1.49), worsening transportation (d = 0.91), and the unequal distribution of tourism's benefits (d = 0.50) — while the two cities did not differ on attitudinal or behavioral items such as observing inappropriate tourist behavior, friction with tourists, loss of local atmosphere, or the broad economic benefits of tourism. This pattern is precisely what social exchange theory would predict. Ap (1992) models resident attitudes as the outcome of an evaluation of the benefits and costs of the tourism exchange, and Andereck, Valentine, Knopf and Vogt (2005) find that benefit-recognition and cost-recognition are separable dimensions, with residents able to perceive strong positives and strong negatives simultaneously. The Kyoto sample fits the "realist" or "ambivalent" resident profile identified across the segmentation literature (Brida, Osti, & Barquet, 2010; Davis, Allen, & Cosenza, 1988; Fredline & Faulkner, 2000; Madrigal, 1995): residents who fully acknowledge tourism's benefits while feeling its material costs acutely. The simultaneous finding that Kyoto residents perceived both a higher upside (recognition and attractiveness, d = 0.54) and much higher costs is a clear example of this high-benefit, high-burden exchange.

Why these particular costs, and why Kyoto, is best explained by the contrasting urban geographies of the two cities. Kyoto is a compact heritage city in which tourism is funnelled into a few honeypot districts — Gion, Higashiyama and Kiyomizu-dera, Arashiyama, Fushimi Inari — along narrow streets within a historic grid, served by a bus network that residents depend on for commuting and schooling. In 2024 the city recorded a record high of 10.88 million foreign visitors, an increasing share of whom stayed overnight (Kyoto City, 2025). In this setting the material costs measured in this study are not abstractions but documented daily realities. Public transportation crowding is among the concerns most frequently reported by residents, with 67.0% troubled by crowded buses and trains and 71.4% by crowded tourist areas (Kyoto City, 2025a); in response, the city has introduced limited-express sightseeing bus services that concentrate visitors on direct routes to major attractions, easing pressure on the local lines that residents depend on (Kyoto City Transportation Bureau, n.d.). The conversion of machiya townhouses and residential buildings into hotels and short-term rentals has been linked to tourism-driven gentrification and the displacement of residents from the historic core; the city's most recent survey recorded an average loss of approximately 700 machiya per year, with the surviving stock falling from 40,146 to 34,580 between 2016 and 2024 (Kyoto City Urban Planning Bureau, 2025), a process shown to have contributed to population decline in central Kyoto (Matsushita, 2023). The commercialization of districts such as Gion has likewise been visible enough to prompt formal restrictions on photography and on tourist access to private alleys, measures aimed at protecting residents' daily lives (Kyoto City, 2025a). The same concentration that produces crowding may also make tourism-linked price pressures more visible to residents. Broader macroeconomic conditions, including the weak yen and higher inbound demand, may have intensified residents' perceptions that tourism contributes to local price pressures, which may help explain why rising living costs produced the single largest effect in

the study. Tokyo, by contrast, is a vast polycentric megacity that absorbs and diffuses tourism across many districts, so that its material and cost impacts are diluted within a far larger and more diversified economy. Where Tokyo experiences tourism friction, it tends to be concentrated in time and place — such as the Shibuya Scramble Crossing, or crowd-control measures around major seasonal events — rather than as the pervasive cost-of-living strain reported in Kyoto. The divergence between the two cities, in other words, is not an artifact of differing attitudes toward tourists but a direct consequence of how each city's spatial structure mediates the same national tourism boom.

These results fit well with both of the major theoretical traditions in resident-attitude research. Through the lens of social exchange theory, Kyoto residents are engaged in a less favorable exchange: they continue to recognize tourism's benefits, but the rising material costs tilt the cost–benefit balance in a way that erodes their sense that current levels are acceptable, which in turn connects back to the RQ1 threshold finding. Through the lens of Doxey's Irrindex, the concentration of tangible burdens in Kyoto is consistent with a host community advancing toward the later, more antagonistic stages of the saturation sequence. That the divergence appears on material rather than attitudinal items suggests that the mechanism most strongly associated with Kyoto's stronger sense of overtourism is economic and infrastructural pressure on everyday life, rather than hostility toward tourists as people — a distinction with important implications for how the problem should be addressed.

RQ4: Policy attitudes and where current governance falls short

The policy findings reinforce this interpretation. The clearest city difference in RQ4 was Kyoto residents' stronger support for public transport and infrastructure improvements ($d = 0.54$), which maps directly onto the transport strain they reported in RQ3 and aligns with the transport-focused pillar of Japan's national overtourism countermeasure package (Japan Tourism Agency, 2023). Notably, there were no robust city differences on limiting tourist numbers or price-based instruments such as accommodation taxes and entry fees, while dispersing tourists to less crowded areas showed only weak and non-robust evidence of a Kyoto–Tokyo difference — even though Kyoto residents experience tourism's costs more acutely. This suggests that residents' response to material burden is a desire to expand capacity and improve daily functioning, rather than to suppress tourism outright. It also speaks to the limits of the instruments most prominent in current policy. Kyoto's accommodation tax, which from March 2026 has risen to a maximum of ¥10,000 per night — a tenfold increase and the highest in Japan (Kyoto City Tourism Association, 2026) — is best understood as a revenue-raising and signalling instrument rather than a demand-management tool. This is consistent with the broader empirical literature on hotel and accommodation taxes, which generally finds negligible effects on visitor demand or hotel revenue even after significant rate increases (Bonham & Gangnes, 1996; Nepal & Nepal, 2019). The absence of a city difference on price-based measures in this study is consistent with residents perceiving that these instruments do not directly relieve the everyday pressures they feel.

Perhaps most telling is what did not differ: residents of both cities expressed similarly modest levels of trust in current tourism governance (RQ4F, $M_{Tokyo} = 2.83$, $M_{Kyoto} = 2.59$) and similarly limited confidence that existing policies are working (RQ4E, $M_{Tokyo} = 2.56$, $M_{Kyoto} = 2.31$). Given that both city means fell below the neutral midpoint of the five-point scale, this shared skepticism, even in Tokyo where material burdens are lower, suggests that dissatisfaction with tourism management is not unique to the most heavily affected city but is a broader feature of how residents view current governance. The political-trust literature holds that trust is a precondition for residents' support of tourism policy, and that low trust may undermine the legitimacy and effectiveness of management efforts (Nunkoo & Gursoy, 2017). The implication is that revenue and infrastructure measures alone are unlikely to be sufficient; building resident trust through transparent, participatory planning — a central recommendation of the UNWTO (2018) framework — may be as important as any single fiscal or physical intervention. Where current governance appears to fall short, then, is less in the absence of tools than in the gap between the instruments being deployed (taxes, manner campaigns) and the material, place-specific burdens that residents actually report.

Conclusion

This study shows that overtourism in Tokyo and Kyoto is not best understood as a fixed visitor count. Residence in both cities associated overtourism more strongly with concentration of tourists in particular places and times than aggregate visitor numbers. Nevertheless, Kyoto residents were more likely to believe that tourism had already exceeded an acceptable number. The clearest differences between the two cities concerned transport disruption, rising living costs, commercialization of cultural spaces, and unequal benefit distribution rather than tourist behaviour and general resident-tourist friction.

Theoretical implications

Beyond the two cities studied, these findings contribute to the broader debate on how overtourism should be conceptualized. They support a perception-based, context-sensitive definition of the overtourism threshold: rather than tracking absolute visitor numbers, the threshold appears to be shaped by residents' lived experience of tourism's material costs. Kyoto and Tokyo are exposed to the same national tourism boom, yet residents diverge sharply in judging whether that boom has become excessive — a divergence that tracks material burden rather than attitudes toward tourists themselves. This is consistent with Koens, Postma and Papp's (2018) argument that overtourism is a localized, multidimensional phenomenon rather than a function of aggregate numbers, and with the UNWTO's (2018) emphasis on resident perceptions as central to defining sustainable urban tourism growth.

The findings also refine the application of social exchange theory to overtourism. Rather than a single net balance of costs against benefits (Ap, 1992), benefit-recognition and cost-recognition operate as separable dimensions (Andereck et al., 2005): Kyoto residents simultaneously registered stronger recognition of tourism's benefits and substantially stronger

recognition of its costs. The threshold for judging tourism unacceptable is therefore reached not when residents cease to value it, but when its material burdens accumulate independently of that value, including a sense of inequity in how tourism's benefits are shared (RQ3J), which is conceptually distinct from cost items tied to price and commercialization. This conceptual distinction between burden and inequity extends existing accounts of the benefit–cost structure of resident attitudes and suggests that models of overtourism should treat distributive fairness as a dimension distinct from cost magnitude.

Practical implications

These findings point toward management strategies tailored to urban form. The two cities' current policy trajectories suggest both the promise and the limits of this approach. In compact heritage cities such as Kyoto, where tourism concentrates in a few honeypot districts and material costs are tightly coupled to that concentration, the priority is relieving the specific pressures residents identify. Kyoto's recent measures move in this direction. These include limited-express sightseeing buses that separate visitor flows from the local lines residents depend on (Kyoto City Transportation Bureau, n.d.), a planned citizen-priority bus fare scheduled for introduction in 2027, and an accommodation tax whose revenue is earmarked for infrastructure and overtourism countermeasures (Kyoto City Industry and Tourism Bureau, 2025). First, because the largest perceived material burden in this study concerned rising prices and living costs (RQ3I) rather than transport alone, revenue could be directed toward measures that directly ease everyday cost-of-living pressures on residents, rather than concentrating on visitor-facing infrastructure. Second, dispersal-based measures common to both cities warrant caution: soft interventions encouraging visitors toward less-crowded areas have shown limited effect on behavior, since demand remains concentrated on iconic sites (Zuckerman Farkash et al., 2023). Resident-oriented transport measures, such as citizen-priority fares where implemented or proposed, are likely to be more tractable than redistributing demand.

The finding that perceived inequity in benefit-sharing (RQ3J) constitutes a burden distinct from cost items also points to a more targeted mechanism than uniform city-wide spending. Protected-area tourism has used proximity-based revenue-sharing, returning a share of visitor-generated revenue directly to the communities most exposed to tourism's pressures, with documented success in building resident support (Snyman et al., 2023). The present study suggests that an urban analogue — directing a defined share of accommodation-tax revenue generated within specific districts toward visible, local benefits for residents in the immediately surrounding zones — would adapt this logic to a compact heritage city. Because this study's sampling was itself structured around residents' proximity to tourist hotspots, it offers one template for identifying which residents such a scheme should prioritize.

In polycentric Tokyo, where tourism friction is diluted across a larger economy and surfaces mainly as localized, time-bounded episodes, place- and event-specific management is more proportionate, as in Shibuya Ward's ordinances restricting street drinking around Halloween and New Year (Shibuya Ward, 2024). Because the burden here is episodic rather than

pervasive, the risk is not under-provision but the gradual conversion of temporary, event-specific controls into standing restrictions on public space; proportionate management means keeping interventions bound to the times and places where friction is demonstrated.

Finally, the shared and modest distrust of current governance found in both cities has its own weight: residents in Tokyo expressed skepticism about policy effectiveness and authorities comparable to that in more heavily burdened Kyoto, indicating that credibility cannot rest on instrument choice alone. Trust in governing institutions is an important factor shaping residents' support of tourism policy, and low trust can undermine the legitimacy and effectiveness of management efforts (Bramwell, 2011; Nunkoo & Gursoy, 2017). Pairing burden-relief measures with transparent reporting on their effects — a practice consistent with the participatory planning the UNWTO (2018) framework recommends — would allow residents to assess whether the instruments deployed are responsive to the burdens they themselves report.

Several limitations point to directions for future research. First, although the present sample (N = 101) was sufficient to detect the large city-level differences reported here, larger and more representative samples would allow these patterns to be generalized with greater confidence. Ideally, such samples would be stratified by age, length of residence, and proximity to tourist hotspots, which would also support more robust modeling of within-city variation. Second, this study compared residents at the city level, treating Kyoto and Tokyo as whole units, and for the proximity measure it anchored each city to a single representative hotspot (Shibuya Scramble Crossing in Tokyo, and Gion and Kiyomizu-dera in Kyoto). Yet tourism pressure in both cities is distributed across multiple hotspots rather than a single site. Tokyo also concentrates visitors in districts such as Asakusa (Taito), Shinjuku, and Ueno, while Kyoto's pressure extends to Arashiyama and Higashiyama beyond the Gion area. District-level research that distinguishes these heavily affected neighborhoods from less-exposed areas would therefore clarify how far the observed city differences are associated with specific hotspots rather than with the city as a whole, and it would capture within-city variation that a single reference point necessarily overlooks. Third, the cross-sectional design captures perceptions at a single point in time. Longitudinal research would show whether residents' thresholds and tolerance shift as visitor numbers, prices, and policies continue to change, and whether Doxey's proposed progression from apathy toward irritation and antagonism actually unfolds over time. Finally, this study relied on survey self-reports alone. Mixed-methods designs that combine surveys with interviews, observation, and objective data, such as housing-market and short-term-rental records or mobility and congestion data, would triangulate residents' perceptions against material conditions on the ground and strengthen causal interpretation of how concentration translates into perceived burden. Finally, this study examined residents only. The UNWTO (2018) definition of overtourism encompasses both residents' quality of life and visitors' experience, so a full account of when tourism becomes excessive would also need to incorporate the visitor side, which this resident-focused design does not capture.

Works Cited

Andereck, K. L., Valentine, K. M., Knopf, R. C., & Vogt, C. A. (2005). Residents' perceptions of community tourism impacts. Annals of Tourism Research, 32(4), 1056–1076.

Ap, J. (1992). Residents' perceptions on tourism impacts. Annals of Tourism Research, 19(4), 665–690.

Bonham, C. S., & Gangnes, B. (1996). Intervention analysis with cointegrated time series: The case of the Hawaii hotel room tax. Applied Economics, 28(10), 1281–1293. https://doi.org/10.1080/000368496327831

Bramwell, B. (2011). Governance, the state and sustainable tourism: A political economy approach. Journal of Sustainable Tourism, 19(4–5), 459–477.

Brida, J. G., Osti, L., & Barquet, A. (2010). Segmenting resident perceptions towards tourism — a cluster analysis with a multinomial logit model of a mountain community. International Journal of Tourism Research, 12(5), 591–602.

Burtis, C., & Wise, N. (2024). Viewpoint: The overtourism strain on Kyoto: Economic policies and community impacts. Local Economy, 39(5–6), 283–288. https://doi.org/10.1177/02690942251350117

Clawson, M., & Knetsch, J. L. (1966). Economics of Outdoor Recreation. Johns Hopkins University Press.

Davis, D., Allen, J., & Cosenza, R. M. (1988). Segmenting local residents by their attitudes, interests, and opinions toward tourism. Journal of Travel Research, 27(2), 2–8.

Doxey, G. V. (1975). A causation theory of visitor–resident irritants: Methodology and research inferences. In Proceedings of the Travel Research Association 6th Annual Conference (pp. 195–198). San Diego, CA.

Fredline, E., & Faulkner, B. (2000). Host community reactions: A cluster analysis. Annals of Tourism Research, 27(3), 763–784.

Goodwin, H. (2017). The challenge of overtourism (Responsible Tourism Partnership Working Paper 4).

Handler, I. (2026). Overtourism in Kyoto: Resident perceptions, emotional distress, and governance failure. Cities, 173, Article 106918. https://doi.org/10.1016/j.cities.2026.106918

Japan National Tourism Organization (JNTO). (2026). Japan tourism statistics: Visitor arrivals. https://statistics.jnto.go.jp/en/

Japan Tourism Agency. (2023). Package of countermeasures against overtourism. Ministry of Land, Infrastructure, Transport and Tourism.

Japan Tourism Agency. (2024). White paper on tourism in Japan, 2024 (Summary). Ministry of Land, Infrastructure, Transport and Tourism.

Kato, H. (2025). Spatial patterns and geographic characteristics of tourism-accommodation intensity hotspots in Kyoto city. Annals of Tourism Research Empirical Insights, 6(1), 100178.

Koens, K., Postma, A., & Papp, B. (2018). Is overtourism overused? Understanding the impact of tourism in a city context. *Sustainability, 10*(12), 4384.

Kyoto City. (2021). *Kyoto Tourism Promotion Plan 2025*.

Kyoto City. (2024). *Kyoto tourism-related statistics*.

Kyoto City. (2025). *Results of the 2024 Kyoto Tourism Comprehensive Survey*.

Kyoto City Industry and Tourism Bureau, Tourism and MICE Promotion Office. (2025, October). *Kyoto Tourism and MICE Promotion Plan 2030 (tentative name): Interim draft* [Public comment document]. https://www.city.kyoto.lg.jp/templates/pubcomment/cmsfiles/contents/0000346/346394/pabliccommentgaiyoueng.pdf

Kyoto City Tourism Association (DMO KYOTO). (2026). Kyoto's accommodation tax to change starting March. *Kyoto Travel*. https://kyoto.travel/en/news/tax-change/

Kyoto City Transportation Bureau. (n.d.). Sightseeing limited express bus. *Kyoto City Bus & Subway Information Guide*. https://www2.city.kyoto.lg.jp/kotsu/webguide/en/bus/limited_express.html

Kyoto City Urban Planning Bureau. (2025, December 25). *Reiwa 6 Kyomachiya status survey* [令和6年度京町家状況調査]. https://www.city.kyoto.lg.jp/tokei/page/0000348982.html

Madrigal, R. (1995). Residents' perceptions and the role of government. *Annals of Tourism Research, 22*(1), 86–102.

Matsushita, D. (2023). Population decline and urban transformation by tourism gentrification in Kyoto City. *Sustainability, 15*(3), 2247. https://doi.org/10.3390/su15032247

Mihalič, T. (2020). Conceptualising overtourism: A sustainability approach. *Annals of Tourism Research, 84*, 103025.

Muler Gonzalez, V., Coromina, L., & Galí, N. (2018). Overtourism: Residents' perceptions of tourism impact as an indicator of resident social carrying capacity — case study of a Spanish heritage town. *Tourism Review, 73*(3), 277–296.

Nepal, R., & Nepal, S. K. (2019). Managing overtourism through economic taxation: Policy lessons from five countries. *Tourism Geographies, 23*(5–6), 1094–1115.

Nunkoo, R., & Gursoy, D. (2017). Political trust and residents' support for alternative and mass tourism: An improved structural model. *Tourism Geographies, 19*(3), 318–339.

O'Reilly, A. M. (1986). Tourism carrying capacity: Concept and issues. *Tourism Management, 7*(4), 254–258.

Sharpley, R. (2014). Host perceptions of tourism: A review of the research. *Tourism Management, 42*, 37–49.

Shibuya Ward. (2024). *Ordinance to prevent nuisance behaviour in public spaces* [Revised ordinance on street drinking]. Shibuya City.

Snyman, S., Fitzgerald, K., Bakteeva, A., Ngoga, T., & Mugabukomeye, B. (2023). Benefit-sharing from protected area tourism: A 15-year review of the Rwanda tourism revenue sharing programme. *Frontiers in Sustainable Tourism, 1*, 1052052. https://doi.org/10.3389/frsut.2022.1052052

Tokyo Convention & Visitors Bureau. (2016). *Tokyo tourism statistics: Number of foreign visitors to Tokyo*. Tokyo Convention & Visitors Bureau.

Tokyo Convention & Visitors Bureau. (n.d.). *Responsible travel* [GO TOKYO]. https://www.gotokyo.org/en/

Tokyo Metropolitan Government. (2025). *Tokyo tourism data catalog*. https://data.tourism.metro.tokyo.lg.jp/en/

World Tourism Organization (UNWTO). (2018). *'Overtourism'? Understanding and managing urban tourism growth beyond perceptions*. Madrid: UNWTO.

Zuckerman Farkash, M., Birenboim, A., Fleischer, A., & Ben-Nun Bloom, P. (2023). Can local tours disperse tourists from city centres? *Current Issues in Tourism, 27*(4), 511–516. https://doi.org/10.1080/13683500.2023.2218607

Stem Cell-Based Therapeutics in Amyotrophic Lateral Sclerosis: Therapeutic Potential and Safety Challenges By Koji Karasawa

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurological disorder that causes continuous degeneration of the patient's motor neurons. In the initial stages, patients experience symptoms such as muscle weakness and progressive decline in motor function. As the disease advances, symptoms inevitably worsen into complete paralysis and respiratory failure, resulting in death. Although the precise etiology of ALS is unclear, several mechanisms are attributed to its pathophysiology. While gene mutations are often the main catalyst of familial ALS, sporadic ALS is multifactorial, driven by a combination of pathological and environmental factors. Because of the disease's intricate pathophysiology, currently approved treatments cannot cure ALS. Moreover, these treatments cannot address more than one specific pathological mechanism. These limitations necessitate a treatment approach that can simultaneously address the numerous pathological mechanisms of ALS. Stem cells are cells that can differentiate into various specialized cell types and continuously proliferate. Stem cells also have unique properties, including anti-inflammatory, immunomodulatory, and neuroprotective effects, as well as the ability to recapitulate patient-specific phenotypes. Stem cells can respond and adapt to disease-induced biochemical reactions. Such characteristics make stem cells ideal for combating the multifactorial pathomechanisms underlying ALS. This review primarily provides an overview of different types of stem cells, including ESCs, iPSCs, NSCs, and MSCs, and their therapeutic applications in ALS. Although each stem cell type has distinct safety and ethical limitations, promising results in mouse and patient studies, as well as efforts to address these limitations, ultimately validate stem cell therapy as a prospective approach to ameliorate ALS. For further success in the development of stem cell-based therapeutics in ALS, rigorous clinical trials and safety assessments must be implemented.

Introduction

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a progressive neurodegenerative disease that induces the continuous depletion of motor neurons (Papadeas and Maragakis). ALS can take on both sporadic and familial disease forms, which essentially means it can emerge spontaneously or be inherited genetically. The US prevalence of ALS from estimated statistics of 2022 shows around 32,893 cases, which is a prevalence of approximately 9.9 people per 100,000. Trend analysis of existing data also indicates a continuous increase through 2030, culminating in 10.5 people per 100,000, or 36,308 cases (Mehta et al.). ALS causes patients to experience symptoms such as muscle weakness and twitching in early stages, eventually deteriorating to loss of respiratory function and inability to perform essential daily actions (Brotman et al.; Brown and Al-Chalabi). Currently, there are no permanent cures for ALS; only treatments that hinder the progression of the disease and/or lessen the symptoms, which raises severe concerns for society and patient quality of life (Brotman et al.).

Stem cells have the unique ability to differentiate into specialized cell types depending on their potency (Figure 2). Stem cells can also self-replicate and renew perpetually (Zakrzewski et al.). Stem cell-based treatments are considered a paradigm in regenerative medicine for conditions with irreversible symptoms due to their proliferative, differentiative, and other stem cell type-specific properties. Such conditions include neurodegenerative diseases, blood diseases, autoimmune diseases, cardiovascular diseases, metabolic and endocrine disorders, respiratory diseases, and other tissue degenerative diseases (Aly; Hoang et al.). ALS's consequence of irreversible loss of motor neurons, its multifactorial pathophysiology, the lack of a definitive treatment, and the robust properties of stem cells provide the rationale for using stem cell-based approaches as a treatment method. In this review, I aim to provide an overview of ALS, stem cells, and their existing applications, primarily focusing on the specific uses of stem cells in ALS treatment and the limitations of this approach.

Overview of Amyotrophic Lateral Sclerosis

ALS is the most common motor neuron disease in a group of diseases that affect upper and lower motor neurons and is both familial and sporadic, meaning it can be inherited or manifest randomly. The symptoms of ALS vary according to the severity of the disease: patients in early stages experience muscle weakness and twitching, with symptoms progressing into muscle and respiratory paralysis in later stages, resulting in death (Brotman et al.; Brown and Al-Chalabi; Alves et al.).

The precise cause of ALS is not fully understood; however, possible pathological mechanisms of ALS include gene mutations, proteostasis failure, oxidative stress, mitochondrial dysfunction, neuronal signaling impairment, such as synaptic dysfunction and axonal transport disturbance, neuroinflammation, and excitotoxicity (Brotman et al.; Tzeplaeff et al.; Margeta) (Figure 1). There are also environmental risk factors associated with ALS, which include exposure to heavy metals, pesticides, electromagnetic fields, and neurotoxins from high-risk occupations; military work; tobacco use; and frequent head injuries from sports (Brotman et al.; Brown and Al-Chalabi; Quinn and Elman). Gene mutations are the driving factor behind familial ALS cases, which only make up about 5-10% of all occurrences (Brotman et al.). However, the manifestation of sporadic ALS is a result of genetic mutations, other pathological mechanisms, and environmental factors operating in conjunction (Brotman et al.).

Upwards of 120 genes have been reported to be associated with ALS, and over 20 genes have been identified to be causative in familial ALS (Brotman et al.; Brown and Al-Chalabi; H. Wang et al.). Notable genes include Superoxide Dismutase 1 (*SOD1*), Transactive response DNA binding protein (*TARDBP)*, Fused in Sarcoma (*FUS*), Chromosome 9 Open Reading Frame 72 (*C9orf72*), Optineurin (*OPTN*), Valosin-Containing Protein (*VCP*), Ubiquilin 2 (*UBQLN2*), and TANK-binding kinase 1 (*TBK1*) (Ciervo et al.). However, out of these genes, *SOD1*, *TARDBP*, *FUS*, and *C9orf72* have been found to mutate most often and thus are considered to play the most substantial roles in causing familial ALS (Grassano et al.; Ivantsik et al.).

The SOD1 protein is a free radical scavenger that converts superoxide into oxygen gas and hydrogen peroxide, located in the intermembrane space of the mitochondria and cytosol of the cell (Bunton-Stasyshyn et al.; Paez-Colasante et al.). *SOD1* mutations account for approximately 20% of familial ALS cases, with certain mutations inducing SOD1 misfolding and aggregation exhibiting properties similar to prions that contribute to the pathogenesis of ALS (Paez-Colasante et al.; Turner et al.; Lobsiger et al.). These mutations include: W32S, H46R, G85R, G93A, and G127X (Ivantsik et al.). Misfolded SOD1 with the aforementioned mutations can self-propagate, creating an accumulation of misfolded protein clusters that induce toxic gain and result in ALS (Paez-Colasante et al.; Polymenidou and Cleveland; Ivantsik et al.; Brasil et al.). *SOD1* mutations that engender the onset of ALS are suggested to be those that foster protein misfolding and aggregation through the collapse of the protein's homodimeric complex, thereby impacting its structural integrity (Ivantsik et al.; Brasil et al.).

The TDP-43 protein, derived from the *TARDBP* gene, is an RNA-binding protein. TDP-43 contributes to the process of gene expression: repression and activation of transcription, splicing of RNA, and transport of mRNA from the cell nucleus to the cytoplasm (Zeng et al.; Duan et al.; Paez-Colasante et al.). *TARDBP* has more than 40 distinct non-synonymous mutations associated with the pathogenesis of familial ALS (Paez-Colasante et al.; Narayanan et al.; Honda et al.). Among these *TARDBP* mutations, 22 are associated with the misfolding of TDP-43, including A90V, G290A, G294A, G294V, A315T, A315E, M337V, Q343R, G348C, N352S, and A382T (Ivantsik et al.). However, there is no evidence of protein misfolding directly instigated by specific *TARDBP* mutations (Ivantsik et al.). Most *TARDBP* mutations linked with ALS occur in the C-terminus of TDP-43, the end region of the protein (Paez-Colasante et al.). These regions are rich in glycine and lack proper shape and structure, and are frequently seen in misfolded TDP-43 aggregates of ALS (Paez-Colasante et al.; Meyerowitz et al.; Buratti et al.; Johnson et al.; Sang Hwa Kim et al.). Misfolded TDP-43 aggregates form via the displacement of irregularly phosphorylated, ubiquitin-conjugated, and proteolytically processed proteins in the cell cytoplasm (Ivantsik et al.). Such aggregates, implicated in ALS pathogenesis, inhibit the RNA-binding function of TDP-43, resulting in the disruption of the RNA life cycle and gene expression process (X. Wang et al.).

The FUS protein, like TDP-43, is also an RNA-binding protein that is involved in gene expression. Located on chromosome 16, mutations in *FUS* affect the C-terminal of FUS (Paez-Colasante et al.). The *C9orf72* gene is a part of the 9p21 region of chromosome 9, and the C9orf72 protein is a guanine nucleotide exchange factor for Rho-GTPases and Rab-GTPases (Iyer et al.). A specific repeat expansion mutation of the GGGGCC hexanucleotide in intron 1 of *C9orf72*, passed down through autosomal dominant inheritance, is considered the most common genetic cause of ALS (Paez-Colasante et al.; Renton et al.; DeJesus-Hernandez et al.).

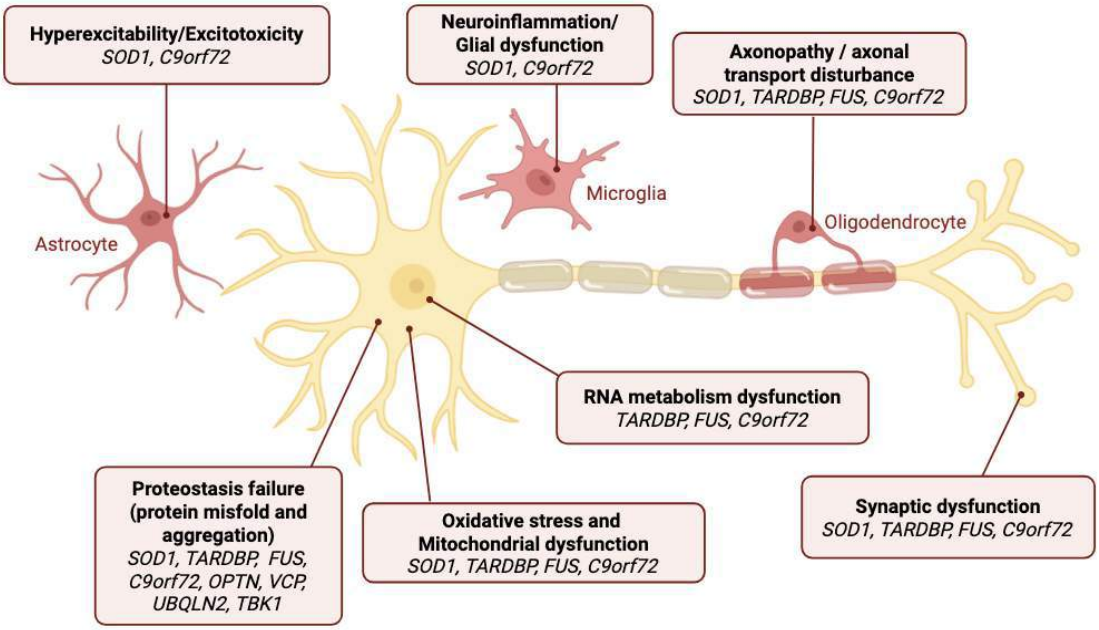


Figure 1: Major pathophysiological mechanisms and key molecules in ALS. ALS has several pathophysiological mechanisms, including hyperexcitability, excitotoxicity, neuroinflammation, glial dysfunction, axonopathy, proteostasis failure, oxidative stress/mitochondrial dysfunction, RNA metabolism dysfunction, and synaptic dysfunction. Different gene mutations are associated with each mechanism. Abbreviations: Superoxide Dismutase 1 (*SOD1*); Transactive response DNA binding protein (*TARDBP*); Fused in Sarcoma (*FUS*); Chromosome 9 Open Reading Frame 72 (*C9orf72*); Optineurin (*OPTN*); Valosin-Containing Protein (*VCP*); Ubiquilin 2 (*UBQLN2*); TANK-binding kinase 1 (*TBK1*).

Overview of Stem Cell Types and Stem Cell Therapy

Embryonic stem cells (ESCs)

Originating from the inner cell mass of the pre-implantation blastodermic vesicle, embryonic stem cells (ESCs) were obtained using procedures and methods that enabled the separation of embryonic carcinoma cells (ECCs) from tumors (Rippon and Bishop; Evans and Kaufman; Martin). Within the pre-implantation blastodermic vesicles are the inner cell mass and trophoblast cells that give rise to the placenta. Inner cell mass cells give rise to embryos, which have the ability to become all specialized cell types. ESCs are pluripotent stem cells, meaning they can differentiate into every cell type except the placenta. ESCs can also be maintained in a state of perpetual proliferation, making them self-renewing.

Murine ESCs specifically enabled research into the preliminary stages of mammalian development, understanding of pluripotency and cell specialization mechanisms, and contributed to research in gene regulation and expression in recombinant animals through the development of genome-modifying technology (Rippon and Bishop; Thomas and Capecchi).

Induced pluripotent stem cells (iPSCs)

Induced pluripotent stem cells (iPSCs) are cells that regain pluripotency through the reprogramming of specialized mature cells, giving them properties that replicate ESCs (Lee et al.; Zakrzewski et al.; Chehelgerdi et al.; Choi et al.). iPSCs have the capacity to become any cell type other than the placenta and can endlessly self-replicate, making them extremely useful for regenerative medicine and other fields of study. Human iPSCs (hiPSCs) mitigate the ethical considerations of using hESCs, which necessitate the destruction of human embryos in their method of obtaining, making hiPSCs an extremely valuable alternative for use in research and medicine (Singh et al.).

The mechanisms behind hiPSCs enable the creation of patient-specific models and treatments for diseases that have no conventional treatment (Cerneckis et al.; Singh et al.). hiPSCs' cell differentiation and self-propagating properties allow for greater precision in human disease modeling compared to previous methods, such as using animals, including monkeys and other primates; dogs; rats; and mice (Singh et al.). The properties of hiPSCs also allow meticulous understanding in the discovery and testing of different treatments for human diseases (Cerneckis et al.; Singh et al.). Other contemporary applications of hiPSCs include human development modeling, specialized cell transplantation, and gene therapy for diseases (Cerneckis et al.; Lee et al.; Liou et al.).

Neural stem cells (NSCs)

Neural stem cells (NSCs), like all stem cells, have the capacity to self-renew. However, NSCs are unique in the way that they can become all of the main cell types found in the central nervous system (CNS). Specifically, NSCs can become neurons and key glial cells such as astrocytes and oligodendrocytes (Shoemaker and Kornblum; Molyneaux et al.). This classifies NSCs at a differentiation level of multipotent, since they can mature into numerous, but not all, specialized cell types. NSCs can be obtained through a multitude of ways, including amplification via induced proliferation and self-renewal; differentiation from pluripotent iPSCs and ESCs; direct extraction from mature or fetal CNS; and reprogramming of non-neural somatic cells (Shoemaker and Kornblum; Zholudeva et al.; J. Kim et al.).

The application of NSCs is crucial for researching the manner in which neurons gain cell specialization and phenotypes, how neuron and glial cell relations impact the CNS, and cellular therapy for neurological trauma (Zholudeva et al.). CNS-derived NSCs have also been used to study function and development in the brain and spinal cord, which has contributed to the procedural tuning of CNS cell type differentiation from iPSC and ESC (Zhao and Moore; Zholudeva et al.).

Mesenchymal stem cells (MSCs)

Mesenchymal stem cells (MSCs) are self-renewing multipotent progenitor stem cells that have the capacity to differentiate into several specialized cell types of all three germ layers, but primarily the mesodermal layer, which includes muscle, fat, bone, and cartilage (Ding et al.). MSCs can be harvested from a variety of organs and tissues in the body, including the heart,

lungs, bone marrow, body fat, endometrial polyps, umbilical cord, placenta, menstrual fluid, and more (Hass et al.; Ding et al.). The advantage of isolating MSCs from tissues linked with birth, which include the placenta, umbilical cord, and cord blood, is the mitigation of invasive procedures such as bone marrow aspiration, because they are easily accessible immediately postpartum (Hass et al.). MSCs derived from neonatal tissue have been reported to have better differentiation potency and self-renewal capacity compared to MSCs extracted from bone marrow (Hass et al.). MSCs also have homing capabilities when migrating, dispersing to become various organs and tissues (Ding et al.; Devine et al., "Mesenchymal Stem Cells Are Capable"; Devine et al., "Mesenchymal Stem Cells Distribute"; Allers et al.; François et al.).

Modern applications of MSCs include transplantation for heart repair in cardiovascular diseases, specifically treating myocardial infarction (Kim and Cho; S. Zhang et al.; Nagaya et al.; Chen et al.; Zeinaloo et al.); treatment of steroid-resistant graft-versus-host disease after bone marrow transplant (Kim and Cho; Le Blanc et al.; Ringdén et al.; K.-H. Wu et al.); cartilage and bone disease; cirrhosis and other liver disease (Kim and Cho; Mohamadnejad et al.; Kharaziha et al.); autoimmune diseases such as Crohn's disease, systemic lupus erythematosus, and rheumatoid arthritis (Kim and Cho; Garcia-Olmo et al.; Sarsenova et al.; Sun et al.); and promising prospects in cancer treatment (Kim and Cho; Shi et al.).

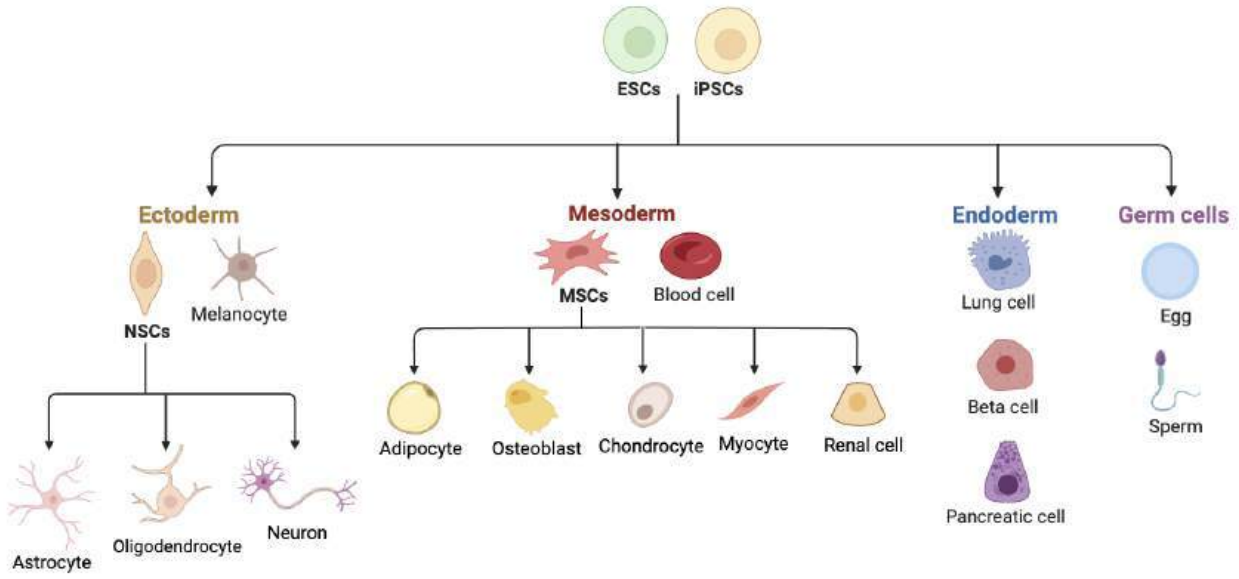


Figure 2: Differentiation potential of major stem cell types. Pluripotent stem cells such as ESCs and iPSCs can differentiate into cells within the ectoderm, mesoderm, and endoderm germ layers, as well as germ cells. NSCs can differentiate into astrocytes, oligodendrocytes, and neurons. MSCs can differentiate into adipocytes, osteoblasts, chondrocytes, myocytes, and in some cases renal cells. Abbreviations: embryonic stem cells (ESCs); induced pluripotent stem cells (iPSCs); neural stem cells (NSCs); mesenchymal stem cells (MSCs).

Applications of stem cells in ALS treatment

ALS is a disease distinguished by the inevitable and continual loss of motor neurons.

Due to the complex and multifaceted nature of ALS pathophysiology, there are no treatments that can cure the disease. Contemporary treatments of ALS, such as riluzole, edaravone, and tofersen, offer a limited extent of benefits. These treatments are only able to target individual underlying mechanisms such as excitotoxicity, oxidative stress, and gene-specific proteostasis failure, but cannot address the extensive pathophysiology of ALS (Tzeplaeff et al.; Brotman et al.). Moreover, such treatments can only slightly hinder disease progression (Tzeplaeff et al.; Brotman et al.). Thus, a more comprehensive and adaptive treatment approach is clinically important.

Beyond stem cells' general proliferative and differentiative qualities, specific stem cell types have additional capabilities, including patient-specific phenotypic disease recapitulation, anti-inflammatory effects, immunomodulation, and neuroprotection (Ichiyanagi et al.; Nagamura-Inoue and He; X. Li et al.). Unlike drugs engineered to deliver specific chemicals, stem cells can provide an adaptive treatment. Stem cells are capable of migration, tissue repair, and releasing secretomes that have the aforementioned effects through paracrine signaling (Mirotsou et al.).

These properties are advantageous for simultaneously tackling the complex pathomechanisms of ALS. The following section outlines four major stem cell types: ESCs, iPSCs, NSCs, and MSCs, addressing their specific applications in ALS therapy (summarized in Table 1), evidence for their therapeutic efficacy, and methodological shortcomings that merit some consideration.

ESCs

The main merits of utilizing ESCs are that they can provide a stable, scalable source for generating neural lineages (Lunn et al.). ESCs are also useful for standardized differentiation protocols and comparative studies across laboratories (Kobold et al.; International Stem Cell Initiative). Within the realm of ALS treatment, the applications of ESCs include disease modeling, high-throughput screening/drug discovery, and cell replacement transplantation (Lunn et al.). Transgenic murine ESC-derived motor neurons were initially used by Di Giorgio et al. (Di Giorgio, Carrasco, et al.) to model and study ALS pathogenesis using the human G93A mutation in the SOD1 gene (Di Giorgio, Carrasco, et al.; de Boer and Eggan). Moreover, Di Giorgio et al. used hESC-derived motor neurons as an ALS disease model, demonstrating the vulnerability of human motor neurons to toxic factors released by astrocytes expressing G93A-mutant SOD1 (Di Giorgio, Boulting, et al.). This finding provided evidence of non-cell-autonomous mechanisms' contribution to motor neuron death in ALS. Murine ESC-derived motor neurons were also used in small-molecule survival screening to identify Kenpaullone, a kinase inhibitor compound that can aid motor neuron survival (Yang et al.). A study by Harper et al. found that motor neuron-committed murine ESCs were capable of axon extension, muscle contraction, and survival after transplantation in mature rats with motor neuron injuries (Harper et al.). Although motor neuron replacement has been proposed as a potential remedy for ALS, other studies have

shown that motor neuron cell replacement only improves movement function and has no effect on ALS disease progression and trajectory (Lunn et al.; López-González et al.).

iPSCs

iPSCs' applications in ALS treatment are vast due to their unique property of being able to be created from specific patients, which enables recapitulation of genetic backgrounds (Ichiyanagi et al.; L. Wang et al.; Amorós et al.). Specifically, iPSCs have been used to model familial ALS patients with SOD1, C9orf72, TARDBP, and FUS gene mutations, enabling the study and research of their contribution to ALS pathogenesis (Du et al.). Disease modeling of sporadic ALS cases is done through the reprogramming of patient fibroblasts into iPSCs, which then transform into motor neurons with the specific pathophysiology of ALS (Du et al.). iPSCs derived from particular patients have been used in disease models for drug/compound screening to identify potential ALS treatment candidates and strategies. This includes Bosutinib, Compound #56, Retigabine, RGFP109, Ropinirole, Mitoxantrone, Niclosamide, and SAHA (Du et al.; Imamura et al.; Wainger et al.; Okano et al.; Tsuburaya et al.; Fang et al.; Kuta et al.; Kato and Sakamoto). Imamura et al. found that Bosutinib, a Src/c-Abl inhibitor, has the capacity to mitigate ALS motor neuron degeneration through increasing autophagy and reducing the amount of misfolded SOD1 proteins in motor neurons. This was performed using high-throughput screening of motor neurons differentiated from iPSCs having an SOD1 mutation (Imamura et al.; Du et al.). Wainger et al. discovered that Retigabine, a Kv7 channel activator, can inhibit hyperexcitability in motor neurons associated with SOD1, C9orf72, and FUS mutations, which prolongs the survival of motor neurons (Wainger et al.; Du et al.). Ropinirole, a non-ergoline dopamine agonist, was identified by Fujimori et al. as an ALS drug candidate using drug screening and specifically multiple-phenotype rescue (Fujimori et al.; Du et al.). Ropinirole was found to inhibit FUS and TDP-43 mislocalization, neurite regression, and stress granule development in ALS patient iPSC-derived motor neurons (Fujimori et al.; Du et al.). More recently, the modulation of interleukin 10 (IL-10) and C-C motif chemokine ligand 2/Monocyte Chemoattractant Protein-1 (CCL2/MCP-1), an anti-inflammatory cytokine and pro-inflammatory chemokine, respectively, has been observed to reduce astrocyte- and microglia-induced neurotoxicity and neuroinflammation in both mutant SOD1 and C9orf72 hiPSC models (Allison and Ebert). The results and findings above demonstrate the vast feasibility of iPSCs for drug screening on patient-specific disease models to identify ALS therapeutic candidates and strategies.

iPSCs have also been used for transplantation or for supporting neural cells after differentiation (Frawley et al.). Risks associated with the aforementioned procedures, such as immune rejection, are reduced due to hiPSCs' capacity to be derived specifically from patients (Frawley et al.). Kondo et al. transplanted neural progenitor cells derived from iPSCs into the spinal cords of mouse models with ALS. These neural progenitor cells further differentiated into astrocytes, extending the mouse model's life expectancy by aiding motor neuron survival (Kondo et al.). Nizzardo et al. transplanted hiPSC-derived ALDHhiSSCloVLA4+ NSCs into the spinal

cords of SOD1G93A mice and observed an increase in lifespan and improvement in neuromuscular function and pathology (Nizzardo, Simone, et al.). In a later study, Nizzardo et al. found that transplanting a different NSC subpopulation, LewisX+CXCR4+β1-integrin+, extended the survival of SOD1G93A mice by protecting motor neurons (Nizzardo, Bucchia, et al.). Specifically, LewisX+CXCR4+β1-integrin+ helped maintain neuromuscular synapse stability, mitigate micro- and macrogliosis, and foster axonal sprouting in the ALS mice. These findings were also established in an in vitro model of human ALS motor neurons cultured with ALS astrocytes (Nizzardo, Bucchia, et al.). The aforementioned results demonstrate the promising potential of hiPSC-based cell therapy approaches for ameliorating ALS pathophysiology.

NSCs

NSCs are less prominent in ALS treatment due to their invasive extraction protocol and ethical considerations compared to iPSCs. However, NSCs have still been used for studies on cell therapy. Srivastava et al. conducted a study where allogeneic firefly luciferase-expressing murine NSCs were transplanted into SOD1G93A mice's cervical spine parenchyma (Srivastava et al.). They observed a significant but short-lived improvement in motor function in the transplant group. However, the transplanted NSCs significantly dwindled in number as the disease progressed. This was attributed to the non-conducive microenvironment within the spinal cord of the mouse model with ALS (Srivastava et al.). There have been more promising results from preclinical studies involving the transplantation of hNSCs in ALS mouse models. Zalfa et al. transplanted hNSCs extracted from fetal brains into the anterior horns of L3-L4 of transgenic ALS rats (Zalfa et al.; Frawley et al.). An increase in survival rate of up to 38% was observed for the treatment group, indicating the potential efficacy of hNSC transplantation as a therapeutic approach (Zalfa et al.; Frawley et al.). Preliminary human safety trials have also been conducted by Glass et al., where they transplanted spinal cord-extracted hNSCs into various locations of the spinal cords of patients with ALS (Glass et al.). This allowed them to conclude that intraspinal transplantation of hNSCs can be safely performed, even at higher dosages (Glass et al.). The results produced by the aforementioned studies demonstrate the potential but limited merit of using NSCs to ameliorate ALS.

MSCs

MSCs are more widely favored in ALS therapy compared to NSCs due to easier, less invasive isolation methods and a stronger safety record (Seung Hyun Kim et al.). MSCs can be derived from various tissues and organs. The three most common origins are umbilical cord, adipose tissue, and bone marrow, in order from least to most invasive extraction procedures (Álvarez-Viejo; Berebichez-Fridman and Montero-Olvera). MSCs have mainly been used for direct transplantation and injections, combined and enhanced cell therapies, and cell-free delivery systems.

Umbilical cord-derived MSCs have the least invasive extraction procedure, mainly obtained from the umbilical cords severed after childbirth. MSCs can be isolated from various areas within and surrounding the umbilical cord through designated procedures. These include: Wharton's jelly, veins, arteries, cord blood, and umbilical amniotic membrane (Nagamura-Inoue and He). Umbilical cord MSCs have been used primarily in direct injections and transplantations as both cell-based and cell-free systems. The main advantages of using umbilical cord MSCs in ALS therapy are their anti-inflammatory, immunomodulatory, and neuroprotective properties (Nagamura-Inoue and He). Tang et al. intravenously injected conditioned medium derived from umbilical cord MSCs, a cell-free system, into SOD1G93A mouse models of ALS (J. Tang et al.). They found that continued injections of umbilical MSC-derived conditioned medium improved the survival of mouse models and potentially slowed ALS progression. Moreover, they found that the conditioned medium treatment reduced neuroinflammation and provided neuroprotection by reducing inflammatory responses in mutated microglia and mouse spinal cords and inhibiting neurotoxic astrocyte conversion (J. Tang et al.). In another study, Barczewska et al. intrathecally injected Wharton's jelly MSCs in 67 patients with ALS (Barczewska et al.). They concluded that the treatment prolonged median patient survival by almost double across all levels of ALS progression severity. Statistically significant evidence was also found for a reduction in the rate of ALS progression in patients post-treatment (Barczewska et al.). The presented findings from these two studies demonstrate the vast potential of umbilical cord-derived MSCs in ameliorating ALS.

The main merits of MSCs derived from adipose tissue are their abundance, minimally invasive isolation procedure, high proliferation, wide range of differentiation, and secretomes (Ciervo et al.; Frawley et al.). The exosomes found in adipose MSC secretomes, extracellular vesicles that are components of the secretomes, allow for the use of cell-free therapies and delivery systems in ALS therapy (Frawley et al.). In a study conducted by Bonafede et al., the effects of adipose MSC exosomes on ALS were tested and compared between intravenous injection and intranasal administration in mouse models of SOD1G93A ALS (Bonafede et al.). They found that repeated doses of adipose MSC exosomes reduced neuroinflammation and protected muscle tissue, lower motor neurons, and neuromuscular synapses in the mouse models of ALS. Overall, adipose MSC exosomes helped improve motor function in the short term (Bonafede et al.). Adipose MSCs have also been used in direct transplantations/infusions, as demonstrated by a case report by Shigematsu et al. (Shigematsu et al.). They reported an increase in Revised Amyotrophic Lateral Sclerosis Functional Rating Scale score, from 43 to 45, after a series of autologous adipose MSC intravenous infusions in a middle-aged male patient with ALS. Furthermore, it was reported that the patient remained well for 7 years post-treatment and more than a decade after experiencing initial symptoms (Shigematsu et al.). Although not a controlled study with a proper sample and control group, Shigematsu et al.'s case report ultimately highlights the potential of adipose MSCs in ALS therapeutics.

MSCs derived from bone marrow have been primarily used for transplantation/injections and enhanced/combined cell therapies. They are notable for their emission of various factors that

foster the survival of motor neurons and form new blood vessels, such as Glial Cell Line-Derived Neurotrophic Factor and Vascular Endothelial Growth Factor (Frawley et al.). Siwek et al. gave patients with sporadic ALS three injections of autologous bone marrow MSCs intrathecally over three months (Siwek et al.; Frawley et al.). They observed a slowing of the disease trajectory in patients with more rapidly progressing ALS, whereas there was no significant change in progression rate for patients with slower progressing ALS. However, there are some limitations in Siwek et al.'s study design: small sample size, non-randomization, no control group, and high attrition rate. The variability in findings and limitations in study design contributed to the lack of statistically significant evidence to support that bone marrow MSC injections slowed the ALS disease course in patients (Siwek et al.; Frawley et al.). Bone marrow MSCs are also commonly used in enhanced and combined cell therapies. Cudkowicz et al. intrathecally injected bone marrow MSCs engineered to increase secretion of neurotrophic factors in patients with ALS (Cudkowicz et al.). Data from the study suggest substantial reductions in neuroinflammation and neurodegeneration, and neurotrophic factor support improvements in patients who received the treatment. Additionally, patients with less severe ALS were observed to maintain more motor and other functions after receiving the treatment. Ultimately, Cudkowicz et al.'s study failed to produce sufficient evidence to support an overall improvement in ALS condition due to an overrepresentation of advanced ALS and disparity in disease severity amongst patients (Cudkowicz et al.).

The aforementioned studies across the four stem cell types have produced encouraging results. However, the majority of those studies are not well-established, systematic clinical trials, thus warranting careful consideration and further investigation into the long-term safety and efficacy of stem cell treatments to ameliorate ALS.

Cell type	Derivation origin	Function	General applications	Applications in ALS	Limitations	References
ESCs	Derived from the inner cell mass of pre-implantation embryos	Pluripotent, able to differentiate into every cell type except the placenta.	Research on early development in mammals Understanding of pluripotency and cell specialization mechanisms Research in gene regulation and expression in transgenic animals	ALS disease modeling Drug discovery/high-throughput screening Differentiated motor neuron transplantation	Ethical: destruction of embryos. Safety: Tumorigenicity (undifferentiated cell contamination + uncontrolled differentiation and proliferation), allogeneic immune rejection.	Volarevic et al.
iPSCs	Reprogramming of specialized mature cells	Pluripotent, able to differentiate	Disease modeling Human	Patient-specific ALS disease modeling	Ethical: informed consent from donors, human reproductive	Cerneckis et al.; Singh et al.; Volarevic et al.

	through insertion of Yamanaka factors.	into every cell type except the placenta.	development modeling Drug research and trialing Cell-based remedies Gene therapy	 Drug/compound screening Studying neuroinflammation Differentiated motor neuron transplantation	cloning, genetic modifications, in vitro gametogenesis, human-animal chimeras. Legal: donor privacy, donor benefits, intellectual property, and patents. Safety: genomic instability, tumorigenicity (undifferentiated cell contamination + uncontrolled differentiation and proliferation)	
NSCs	Direct isolation from the CNS Differentiation from pluripotent stem cells Direct lineage reprogramming	Multipotent, able to differentiate into all of the main cell types found in the CNS.	Studying CNS function and development Studying cell specialization and phenotypes of neurons	Cell therapy/transplantation	Ethical: source of derivation (embryo destruction, abortion). Safety: Immune rejection, improper migration, tumorigenicity, invasive procedure.	Zhao and Moore; Zholudeva et al.; Y. Tang et al.; Ramos-Zúñiga et al.; Fernandez-Muñoz et al.; Arce et al.
MSCs	Enzymatic digestion Density gradient centrifugation Adherence-based techniques (plastic adherence)	Multipotent, can primarily differentiate into cell types within the mesodermal layer (muscle, fat, bone, and cartilage)	Cardiovascular repair Post-BMT GVHD Remediation of autoimmune diseases Liver disease Progression in cancer treatment	Direct transplantation/injection Enhanced or engineered cell therapy Cell-free therapy	Ethical: donor consent. Safety: Thromboembolism, invasive procedure, undesired cell differentiation, and tumorigenicity.	Kim and Cho; Shi et al.; Mattei et al.; Ding et al.; García-Muñoz and Vives; Gudaviciene et al.; Coppin et al.

Table 1: Various stem cells and their applications in general and in ALS. Abbreviations: embryonic stem cells (ESCs); induced pluripotent stem cells (iPSCs); neural stem cells (NSCs); mesenchymal stem cells (MSCs); central nervous system (CNS); bone marrow transplant (BMT); graft-versus-host disease (GVHD)

Therapeutic implications and limitations

Although the use of stem cells in ALS therapy has shown immense potential, with various studies yielding promising findings, there are also many concerns surrounding stem cell variability. Such concerns include the ethicality and safety of the extraction procedures of respective stem cells, as well as patient safety concerns such as immune incompatibility, tumorigenicity, and thromboembolism (Volarevic et al.; Srivastava et al.; Coppin et al.).

ESCs

The most severe limitations of ESCs and their application in ALS treatment are the regulations and ethical considerations surrounding their acquisition and safety considerations, such as risks in transplantation (Volarevic et al.). Such risks include immune rejection and tumorigenicity (Volarevic et al.).

The moral and ethical dilemma surrounding its derivation and origin is the procedure's eradication of the earliest form of life. ESCs are derived from the inner mass of blastocysts, which are pre-implantation embryos. Different countries have adopted varying stances on this concern, with countries including the United States restricting federal funding for hESC research after August 2001 and Italy prohibiting all research involving hESCs (Volarevic et al.; De Trizio and Brennan).

The pluripotency and differentiation plasticity of ESCs pose a tumorigenic risk in vivo (Bulić-Jakus et al.; Volarevic et al.). Any undifferentiated ESCs after transplantation could lead to the formation of teratomas through uncontrolled differentiation and proliferation, as controlling ESC specialization post-administration is challenging (Volarevic et al.). hESCs' predisposition to developing chromosomal aberrations could shift tumor malignancy and lead to the formation of teratocarcinomas (Bulić-Jakus et al.). Teratomas and teratocarcinomas are tumors that contain tissue from all the ectoderm, mesoderm, and endoderm germ layers (Volarevic et al.; Bulić-Jakus et al.). The difference between these two types of tumors is the presence of undifferentiated embryonic carcinoma cells in the latter (Bulić-Jakus et al.). Hentze et al. tested the tumorigenicity of hESCs in immunodeficient mice (Hentze et al.). The study revealed that while teratoma formation varied under different conditions, it can readily occur under minute hESC concentrations. Furthermore, Hentze et al. found that differentiated hESCs pre-transplantation can still form teratomas post-transplantation (Hentze et al.).

iPSCs

iPSCs are widely considered a superior stem cell lineage compared to ESCs because their derivation procedure mitigates the ethical controversy surrounding the destruction of embryos and the risk of allogeneic immune rejection (Volarevic et al.).

iPSCs face the issue of genomic instability, which refers to alterations in the iPSC genome. Genomic instability of iPSCs can arise from three main sources: mutations engendered by somatic cell reprogramming, antecedent donor cell mutations, and mutations developed from continuous passaging (Yoshihara et al.; Volarevic et al.). Variations of iPSC genetic alterations

include single-nucleotide variations, chromosomal abnormalities, and copy number variations (Yoshihara et al.; K. Zhang et al.). Such genome alterations could compromise the differentiation stability and potency of iPSCs, leading to decreased differentiation capacity, unpredictable specialization results, uncontrolled proliferation, and tumorigenicity (K. Zhang et al.). Liang et al. tested four distinct hiPSC lines and found that hiPSC lines with higher genomic instability also had higher tumorigenicity and clonogenicity, suggesting that genomic instability may increase the risk of tumor formation in hiPSCs (Liang et al.).

iPSCs also face the significant safety risk of undifferentiated cell contamination, which could lead to the formation of tumors (Volarevic et al.). Gutierrez-Aranda et al. reported that teratoma formation in non-obese diabetic/severe combined immune-deficient (NOD/SCID) IL2Rγ−/− mice after injection of undifferentiated hiPSCs was more rapid compared to mice that received hESC grafts (Gutierrez-Aranda et al.). Additionally, they observed that the undifferentiated hiPSCs formed teratomas in every single recipient irrespective of the injection site (Gutierrez-Aranda et al.).

Despite being able to avoid the ethical issues around embryos, iPSCs encounter a plethora of other ethical controversies. These include: human reproductive cloning, genetic modifications, in vitro gametogenesis, and human-animal chimeras (Moradi et al.).

NSCs

The main ethical concern surrounding NSCs pertains to how they are obtained. While NSCs can be extracted directly from mature CNS, they can also be acquired from embryonic/fetal CNS and from pluripotent stem cell differentiation (Ramos-Zúñiga et al.). Mature CNS NSCs can be acquired from autopsies, biopsies, and cerebrospinal fluid, but such procedures require donor consent and pose significant safety risks due to invasiveness (Fernandez-Muñoz et al.; Arce et al.). Differentiating NSCs from ESCs faces the prominent ethical issue of embryo destruction (Ramos-Zúñiga et al.). Additionally, extracting NSCs from embryonic CNS is often done with the use of aborted embryos and/or fetuses, which sparks ethical controversy across various religions and countries (Fernandez-Muñoz et al.).

One of the most prominent issues that NSCs face is immune rejection. Immune rejection arises when allogeneic sources of NSCs are transplanted and can occur through either innate or adaptive host immune responses (Srivastava et al.). Ma et al. demonstrate that allogeneic and xenogeneic NSCs transplanted in mice recruit various immune cells, indicative of an immune rejection response mediated by both innate and adaptive pathways (Ma et al.). Additionally, they observed a significant reduction in cell abundance in both NSC types after transplantation (Ma et al.). While the issue of immune rejection can be and is often alleviated with the use of immunosuppressive drugs, they carry their own set of patient health risks if subjected to prolonged exposure. These include hypertension, immunodeficiency, nephrotoxicity, and hepatotoxicity (Master et al.).

MSCs

While the nature of how MSCs are derived avoids pressing ethical controversies such as embryo destruction and abortion faced by other stem cell types, certain methods of MSC extraction encounter patient safety risks, namely procedural invasiveness. The bone marrow required for the derivation of MSCs is extracted using aspiration, which is an invasive and painful procedure. Bone marrow aspiration also comes with the risk of infections, hemorrhage, and traumatic bone lesions (García-Muñoz and Vives; Gudleviciene et al.).

The most immediate and severe patient safety issue associated with transplantation of MSCs is thromboembolism. Upon transplantation and contact with blood, tissue factor on the MSC membrane induces initial clotting (Coppin et al.). This reaction is further intensified by a complementary coagulation response called Instant Blood-Mediated Inflammatory Reaction, ultimately leading to thromboembolism (Coppin et al.). The severity of risk to patient safety is illustrated by a case report by Jung et al., who observed pulmonary embolisms in three patients of the same family after multiple intravenous injections of adipose MSCs (Jung et al.). These findings are further substantiated by a case report by Wu et al., who reported finding venous thromboembolism in two patients after receiving umbilical cord MSCs intravenously (Z. Wu et al.). The patients had distinct renal conditions: chronic kidney disease and kidney transplantation (Z. Wu et al.).

Despite the ethical and safety concerns mentioned above regarding the various stem cells used in treatment, these limitations do not outright invalidate their therapeutic potential. Active efforts to address the various safety issues and risks associated with these stem cells are ongoing and have produced promising results. Park et al. address the issue of tumorigenicity in ESCs and iPSCs by developing K312, a monoclonal antibody that recognizes and eliminates undifferentiated stem cells in vitro prior to transplantation (Park et al.). Ruiz et al. found that increasing Checkpoint kinase 1 expression and adding nucleosides during iPSC reprogramming separately decrease iPSCs' genomic instability (Ruiz et al.). Li et al. mitigated immune rejection in iPSC-generated NSCs by utilizing gene-editing tools to knock out Class II Major Histocompatibility Complex Transactivator and Beta-2-microglobulin and overexpress macrophage migration inhibitory factor (S. Li et al.). Schriner et al. found in vitro that adding low-molecular-weight heparin can diminish thrombin generation in MSCs of various origins, which is promising for reducing the risk of thromboembolism (Schriner et al.).

Conclusion

The purpose of this review was to provide an overview of ALS, prominent stem cell types and their clinical applications within ALS, and the therapeutic implications of such stem cells. ESCs, iPSCs, NSCs, and MSCs each have their respective advantages and limitations in ALS treatment. ESCs are valuable for ALS modeling and creating more specialized cell lineages. iPSCs are more versatile, as they can be used for patient-specific ALS modeling, drug discovery, and autologous cell replacement. NSCs are predominantly used for cell replacement therapy, with existing studies in mice showing promising results of their ALS-ameliorating efficacy. Finally, MSCs can be used in cell-free therapies, direct transplantation, and enhanced/combined

cell therapies, and have a comparatively stronger safety profile than NSCs. Treatment methods that use MSCs are shown to be effective in increasing ALS patient survival. However, ESCs, iPSCs, NSCs, and MSCs each pose major patient safety concerns, including tumorigenicity, genomic instability, immune rejection, and thromboembolism, which inevitably limit the scope of their medical applications. Nevertheless, studies of stem cell-based treatments for ALS thus far have yielded favorable results in both animal models and human patients. Future success in stem cell therapy for ALS is contingent on conducting thorough safety evaluations, conducting long-term clinical trials, and developing more optimized methods to diminish patient risks.

Works cited

Allers, Carolina, et al. “Dynamic of Distribution of Human Bone Marrow-Derived Mesenchymal Stem Cells after Transplantation into Adult Unconditioned Mice.” Transplantation, vol. 78, no. 4, Aug. 2004, pp. 503–08. PubMed, https://doi.org/10.1097/01.tp.0000128334.93343.b3.

Allison, Reilly L., and Allison D. Ebert. “ALS iPSC-Derived Microglia and Motor Neurons Respond to Astrocyte-Targeted IL-10 and CCL2 Modulation.” Human Molecular Genetics, vol. 33, no. 6, Feb. 2024, pp. 530–42. PubMed, https://doi.org/10.1093/hmg/ddad209.

Álvarez-Viejo, María. “Mesenchymal Stem Cells from Different Sources and Their Derived Exosomes: A Pre-Clinical Perspective.” World Journal of Stem Cells, vol. 12, no. 2, Feb. 2020, pp. 100–09. PubMed, https://doi.org/10.4252/wjsc.v12.i2.100.

Alves, Inês, et al. “Demographic Changes in a Large Motor Neuron Disease Cohort in Portugal: A 27 Year Experience.” Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration, vol. 24, no. 7-8, June 2023, pp. 1–11. PubMed, https://doi.org/10.1080/21678421.2023.2220747.

Aly, Riham Mohamed. “Current State of Stem Cell-Based Therapies: An Overview.” Stem Cell Investigation, vol. 7, 2020, p. 8. PubMed, https://doi.org/10.21037/sci-2020-001.

Amorós, Mariana A., et al. “Motor Neuron-Derived Induced Pluripotent Stem Cells as a Drug Screening Platform for Amyotrophic Lateral Sclerosis.” Frontiers in Cell and Developmental Biology, vol. 10, 2022, p. 962881. PubMed, https://doi.org/10.3389/fcell.2022.962881.

Arce, Víctor M., et al. “Neural Stem Cells and Glioblastoma Stem Cells: Redefining Concepts.” Critical Reviews in Oncology/Hematology, vol. 220, Apr. 2026, p. 105187. PubMed, https://doi.org/10.1016/j.critrevonc.2026.105187.

Barczewska, Monika, et al. “Umbilical Cord Mesenchymal Stem Cells in Amyotrophic Lateral Sclerosis: An Original Study.” Stem Cell Reviews and Reports, vol. 16, no. 5, Oct. 2020, pp. 922–32. PubMed, https://doi.org/10.1007/s12015-020-10016-7.

Berebichez-Fridman, Roberto, and Pablo R. Montero-Olvera. “Sources and Clinical Applications of Mesenchymal Stem Cells: State-of-the-Art Review.” Sultan Qaboos University Medical Journal, vol. 18, no. 3, Aug. 2018, pp. e264–77. PubMed, https://doi.org/10.18295/squmj.2018.18.03.002.

Bonafede, Roberta, et al. “ASC-Exosomes Ameliorate the Disease Progression in SOD1(G93A) Murine Model Underlining Their Potential Therapeutic Use in Human ALS.” International Journal of Molecular Sciences, vol. 21, no. 10, May 2020, p. 3651. PubMed, https://doi.org/10.3390/ijms21103651.

Brasil, Aline de Araújo, et al. “Characterization of the Activity, Aggregation, and Toxicity of Heterodimers of WT and ALS-Associated Mutant Sod1.” Proceedings of the National Academy of Sciences of the United States of America, vol. 116, no. 51, Dec. 2019, pp. 25991–6000. PubMed, https://doi.org/10.1073/pnas.1902483116.

Brotman, Ryan G., et al. “Amyotrophic Lateral Sclerosis.” StatPearls, StatPearls Publishing, 2024. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK556151/.

Brown, Robert H., and Ammar Al-Chalabi. “Amyotrophic Lateral Sclerosis.” The New England Journal of Medicine, vol. 377, no. 2, July 2017, pp. 162–72. PubMed, https://doi.org/10.1056/NEJMra1603471.

Bulić-Jakus, Floriana, et al. “Of Mice and Men: Teratomas and Teratocarcinomas.” Collegium Antropologicum, vol. 30, no. 4, Dec. 2006, pp. 921–24.

Bunton-Stasyshyn, Rosie K. A., et al. “SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology: New and Renascent Themes.” The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry, vol. 21, no. 5, Oct. 2015, pp. 519–29. PubMed, https://doi.org/10.1177/1073858414561795.

Buratti, Emanuele, et al. “TDP-43 Binds Heterogeneous Nuclear Ribonucleoprotein A/B through Its C-Terminal Tail: An Important Region for the Inhibition of Cystic Fibrosis Transmembrane Conductance Regulator Exon 9 Splicing.” The Journal of Biological Chemistry, vol. 280, no. 45, Nov. 2005, pp. 37572–84. PubMed, https://doi.org/10.1074/jbc.M505557200.

Cerneckis, Jonas, et al. “Induced Pluripotent Stem Cells (iPSCs): Molecular Mechanisms of Induction and Applications.” Signal Transduction and Targeted Therapy, vol. 9, no. 1, Apr. 2024, p. 112. PubMed, https://doi.org/10.1038/s41392-024-01809-0.

Chehelgerdi, Matin, et al. “Exploring the Promising Potential of Induced Pluripotent Stem Cells in Cancer Research and Therapy.” Molecular Cancer, vol. 22, no. 1, Nov. 2023, p. 189. PubMed, https://doi.org/10.1186/s12943-023-01873-0.

Chen, Shaoliang, et al. “Intracoronary Transplantation of Autologous Bone Marrow Mesenchymal Stem Cells for Ischemic Cardiomyopathy Due to Isolated Chronic Occluded Left Anterior Descending Artery.” The Journal of Invasive Cardiology, vol. 18, no. 11, Nov. 2006, pp. 552–56. PubMed.

Choi, Jiho, et al. “A Comparison of Genetically Matched Cell Lines Reveals the Equivalence of Human iPSCs and ESCs.” Nature Biotechnology, vol. 33, no. 11, Nov. 2015, pp. 1173–81. PubMed, https://doi.org/10.1038/nbt.3388.

Ciervo, Yuri, et al. “Advances, Challenges and Future Directions for Stem Cell Therapy in Amyotrophic Lateral Sclerosis.” Molecular Neurodegeneration, vol. 12, no. 1, Nov. 2017, p. 85. PubMed, https://doi.org/10.1186/s13024-017-0227-3.

Coppin, Louise, et al. “Thrombogenic Risk Induced by Intravascular Mesenchymal Stem Cell Therapy: Current Status and Future Perspectives.” Cells, vol. 8, no. 10, Sept. 2019, p. 1160. PubMed, https://doi.org/10.3390/cells8101160.

Cudkowicz, Merit E., et al. “A Randomized Placebo-Controlled Phase 3 Study of Mesenchymal Stem Cells Induced to Secrete High Levels of Neurotrophic Factors in Amyotrophic Lateral Sclerosis.” Muscle & Nerve, vol. 65, no. 3, Mar. 2022, pp. 291–302. PubMed, https://doi.org/10.1002/mus.27472.

de Boer, A. Sophie, and Kevin Eggan. “A Perspective on Stem Cell Modeling of Amyotrophic Lateral Sclerosis.” Cell Cycle [Georgetown, Tex.], vol. 14, no. 23, 2015, pp. 3679–88. PubMed, https://doi.org/10.1080/15384101.2015.1093712.

De Trizio, Ella, and Christopher S. Brennan. “The Business of Human Embryonic Stem Cell Research and an International Analysis of Relevant Laws.” The Journal of Biolaw & Business, vol. 7, no. 4, 2004, pp. 14–22.

DeJesus-Hernandez, Mariely, et al. “Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS.” Neuron, vol. 72, no. 2, Oct. 2011, pp. 245–56. PubMed, https://doi.org/10.1016/j.neuron.2011.09.011.

Devine, S. M., et al. “Mesenchymal Stem Cells Are Capable of Homing to the Bone Marrow of Non-Human Primates Following Systemic Infusion.” Experimental Hematology, vol. 29, no. 2, Feb. 2001, pp. 244–55. PubMed, https://doi.org/10.1016/s0301-472x(00)00635-4.

Devine, Steven M., et al. “Mesenchymal Stem Cells Distribute to a Wide Range of Tissues Following Systemic Infusion into Nonhuman Primates.” Blood, vol. 101, no. 8, Apr. 2003, pp. 2999–3001. PubMed, https://doi.org/10.1182/blood-2002-06-1830.

Di Giorgio, Francesco Paolo, Gabriella L. Boulting, et al. “Human Embryonic Stem Cell-Derived Motor Neurons Are Sensitive to the Toxic Effect of Glial Cells Carrying an ALS-Causing Mutation.” Cell Stem Cell, vol. 3, no. 6, Dec. 2008, pp. 637–48. PubMed, https://doi.org/10.1016/j.stem.2008.09.017.

Di Giorgio, Francesco Paolo, Monica A. Carrasco, et al. “Non-Cell Autonomous Effect of Glia on Motor Neurons in an Embryonic Stem Cell-Based ALS Model.” Nature Neuroscience, vol. 10, no. 5, May 2007, pp. 608–14. PubMed, https://doi.org/10.1038/nn1885.

Ding, Dah-Ching, et al. “Mesenchymal Stem Cells.” Cell Transplantation, vol. 20, no. 1, 2011, pp. 5–14. PubMed, https://doi.org/10.3727/096368910X.

Du, Hongmei, et al. “Induced Pluripotent Stem Cells and Their Applications in Amyotrophic Lateral Sclerosis.” Cells, vol. 12, no. 6, Mar. 2023, p. 971. PubMed, https://doi.org/10.3390/cells12060971.

Duan, Lauren, et al. “Nuclear RNA Binding Regulates TDP-43 Nuclear Localization and Passive Nuclear Export.” Cell Reports, vol. 40, no. 3, July 2022, p. 111106. PubMed, https://doi.org/10.1016/j.celrep.2022.111106.

Evans, M. J., and M. H. Kaufman. “Establishment in Culture of Pluripotential Cells from Mouse Embryos.” Nature, vol. 292, no. 5819, July 1981, pp. 154–56. PubMed, https://doi.org/10.1038/292154a0.

Fang, Mark Y., et al. “Small-Molecule Modulation of TDP-43 Recruitment to Stress Granules Prevents Persistent TDP-43 Accumulation in ALS/FTD.” Neuron, vol. 103, no. 5, Sept. 2019, pp. 802-819.e11. PubMed, https://doi.org/10.1016/j.neuron.2019.05.048.

Fernandez-Muñoz, Beatriz, et al. “Human Neural Stem Cells for Cell-Based Medicinal Products.” Cells, vol. 10, no. 9, Sept. 2021, p. 2377. PubMed, https://doi.org/10.3390/cells10092377.

François, Sabine, et al. “Local Irradiation Not Only Induces Homing of Human Mesenchymal Stem Cells at Exposed Sites but Promotes Their Widespread Engraftment to Multiple Organs: A Study of Their Quantitative Distribution after Irradiation Damage.” Stem Cells [Dayton, Ohio], vol. 24, no. 4, Apr. 2006, pp. 1020–29. PubMed, https://doi.org/10.1634/stemcells.2005-0260.

Frawley, Lauren, et al. “Stem Cell Therapy for the Treatment of Amyotrophic Lateral Sclerosis: Comparison of the Efficacy of Mesenchymal Stem Cells, Neural Stem Cells, and Induced Pluripotent Stem Cells.” Biomedicines, vol. 13, no. 1, Dec. 2024, p. 35. PubMed, https://doi.org/10.3390/biomedicines13010035.

Fujimori, Koki, et al. “Modeling Sporadic ALS in iPSC-Derived Motor Neurons Identifies a Potential Therapeutic Agent.” Nature Medicine, vol. 24, no. 10, Oct. 2018, pp. 1579–89. PubMed, https://doi.org/10.1038/s41591-018-0140-5.

García-Muñoz, Elisabeth, and Joaquim Vives. “Towards the Standardization of Methods of Tissue Processing for the Isolation of Mesenchymal Stromal Cells for Clinical Use.” Cytotechnology, vol. 73, no. 3, June 2021, pp. 513–22. PubMed, https://doi.org/10.1007/s10616-021-00474-3.

Garcia-Olmo, Damian, et al. “Expanded Adipose-Derived Stem Cells for the Treatment of Complex Perianal Fistula: A Phase II Clinical Trial.” Diseases of the Colon and Rectum, vol. 52, no. 1, Jan. 2009, pp. 79–86. PubMed, https://doi.org/10.1007/DCR.0b013e3181973487.

Glass, Jonathan D., et al. “Transplantation of Spinal Cord-Derived Neural Stem Cells for ALS: Analysis of Phase 1 and 2 Trials.” Neurology, vol. 87, no. 4, July 2016, pp. 392–400. PubMed, https://doi.org/10.1212/WNL.0000000000002889.

Grassano, Maurizio, et al. “Systematic Evaluation of Genetic Mutations in ALS: A Population-Based Study.” Journal of Neurology, Neurosurgery, and Psychiatry, vol. 93, no. 11, July 2022, pp. 1190–93. PubMed, https://doi.org/10.1136/jnnp-2022-328931.

Gudleviciene, Zivile, et al. “Quick and Effective Method of Bone Marrow Mesenchymal Stem Cell Extraction.” Open Medicine [Warsaw, Poland], vol. 10, no. 1, 2015, pp. 44–49. PubMed, https://doi.org/10.1515/med-2015-0008.

Gutierrez-Aranda, Ivan, et al. “Human Induced Pluripotent Stem Cells Develop Teratoma More Efficiently and Faster than Human Embryonic Stem Cells Regardless the Site of Injection.” Stem Cells [Dayton, Ohio], vol. 28, no. 9, Sept. 2010, pp. 1568–70. PubMed, https://doi.org/10.1002/stem.471.

Harper, James M., et al. “Axonal Growth of Embryonic Stem Cell-Derived Motoneurons in Vitro and in Motoneuron-Injured Adult Rats.” Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 18, May 2004, pp. 7123–28. PubMed, https://doi.org/10.1073/pnas.0401103101.

Hass, Ralf, et al. “Different Populations and Sources of Human Mesenchymal Stem Cells (MSC): A Comparison of Adult and Neonatal Tissue-Derived MSC.” Cell

Communication and Signaling: CCS, vol. 9, May 2011, p. 12. PubMed, https://doi.org/10.1186/1478-811X-9-12.

Hentze, Hannes, et al. “Teratoma Formation by Human Embryonic Stem Cells: Evaluation of Essential Parameters for Future Safety Studies.” Stem Cell Research, vol. 2, no. 3, May 2009, pp. 198–210. PubMed, https://doi.org/10.1016/j.scr.2009.02.002.

Hoang, Duc M., et al. “Stem Cell-Based Therapy for Human Diseases.” Signal Transduction and Targeted Therapy, vol. 7, no. 1, Aug. 2022, p. 272. PubMed, https://doi.org/10.1038/s41392-022-01134-4.

Honda, Daiyu, et al. “The ALS/FTLD-Related RNA-Binding Proteins TDP-43 and FUS Have Common Downstream RNA Targets in Cortical Neurons.” FEBS Open Bio, vol. 4, 2013, pp. 1–10. PubMed, https://doi.org/10.1016/j.fob.2013.11.001.

Ichiyanagi, Naoki, et al. “Establishment of In Vitro FUS-Associated Familial Amyotrophic Lateral Sclerosis Model Using Human Induced Pluripotent Stem Cells.” Stem Cell Reports, vol. 6, no. 4, Apr. 2016, pp. 496–510. PubMed, https://doi.org/10.1016/j.stemcr.2016.02.011.

Imamura, Keiko, et al. “The Src/c-Abl Pathway Is a Potential Therapeutic Target in Amyotrophic Lateral Sclerosis.” Science Translational Medicine, vol. 9, no. 391, May 2017, p. eaaf3962. PubMed, https://doi.org/10.1126/scitranslmed.aaf3962.

International Stem Cell Initiative. “Characterization of Human Embryonic Stem Cell Lines by the International Stem Cell Initiative.” Nature Biotechnology, vol. 25, no.7, July 2007, pp. 803-16. PubMed, https://doi.org/10.1038/nbt1318.

Ivantsik, Ouliana, et al. “Exploring Protein Misfolding in Amyotrophic Lateral Sclerosis: Structural and Functional Insights.” Biomedicines, vol. 13, no. 5, May 2025, p. 1146. PubMed, https://doi.org/10.3390/biomedicines13051146.

Iyer, Shalini, et al. “C9orf72, a Protein Associated with Amyotrophic Lateral Sclerosis (ALS) Is a Guanine Nucleotide Exchange Factor.” PeerJ, vol. 6, 2018, p. e5815. PubMed, https://doi.org/10.7717/peerj.5815.

Johnson, Brian S., et al. “TDP-43 Is Intrinsically Aggregation-Prone, and Amyotrophic Lateral Sclerosis-Linked Mutations Accelerate Aggregation and Increase Toxicity.” The Journal of Biological Chemistry, vol. 284, no. 30, July 2009, pp. 20329–39. PubMed, https://doi.org/10.1074/jbc.M109.010264.

Jung, Jae Woo, et al. “Familial Occurrence of Pulmonary Embolism after Intravenous, Adipose Tissue-Derived Stem Cell Therapy.” Yonsei Medical Journal, vol. 54, no. 5, Sept. 2013, pp. 1293–96. PubMed, https://doi.org/10.3349/ymj.2013.54.5.1293.

Kato, Yosuke, and Kazuichi Sakamoto. “Niclosamide Affects Intracellular TDP-43 Distribution in Motor Neurons, Activates Mitophagy, and Attenuates Morphological Changes under Stress.” Journal of Bioscience and Bioengineering, vol. 132, no. 6, Dec. 2021, pp. 640–50. PubMed, https://doi.org/10.1016/j.jbiosc.2021.06.015.

Kharaziha, Pedram, et al. “Improvement of Liver Function in Liver Cirrhosis Patients after Autologous Mesenchymal Stem Cell Injection: A Phase I-II Clinical Trial.” European

Journal of Gastroenterology & Hepatology, vol. 21, no. 10, Oct. 2009, pp. 1199–205. PubMed, https://doi.org/10.1097/MEG.0b013e32832a1f6c.

Kim, Janghwan, et al. “Direct Reprogramming of Mouse Fibroblasts to Neural Progenitors.” Proceedings of the National Academy of Sciences of the United States of America, vol. 108, no. 19, May 2011, pp. 7838–43. PubMed, https://doi.org/10.1073/pnas.1103113108.

Kim, Nayoun, and Seok-Goo Cho. “Clinical Applications of Mesenchymal Stem Cells.” The Korean Journal of Internal Medicine, vol. 28, no. 4, July 2013, pp. 387–402. PubMed, https://doi.org/10.3904/kjim.2013.28.4.387.

Kim, Sang Hwa, et al. “Amyotrophic Lateral Sclerosis-Associated Proteins TDP-43 and FUS/TLS Function in a Common Biochemical Complex to Co-Regulate HDAC6 mRNA.” The Journal of Biological Chemistry, vol. 285, no. 44, Oct. 2010, pp. 34097–105. PubMed, https://doi.org/10.1074/jbc.M110.154831.

Kim, Seung Hyun, et al. “Optimal Therapeutic Strategy of Bone Marrow-Originated Autologous Mesenchymal Stromal/Stem Cells for ALS.” Stem Cells Translational Medicine, vol. 13, no. 4, Apr. 2024, pp. 309–16. PubMed, https://doi.org/10.1093/stcltm/szad095.

Kobold, Sabine, et al. “Human Embryonic and Induced Pluripotent Stem Cell Research Trends: Complementation and Diversification of the Field.” Stem Cell Reports, vol. 4, no. 5, 2015, pp. 914-25. PubMed, https://doi.org/10.1016/j.stemcr.2015.03.002.

Kondo, Takayuki, et al. “Focal Transplantation of Human iPSC-Derived Glial-Rich Neural Progenitors Improves Lifespan of ALS Mice.” Stem Cell Reports, vol. 3, no. 2, Aug. 2014, pp. 242–49. PubMed, https://doi.org/10.1016/j.stemcr.2014.05.017.

Kuta, Rachel, et al. “Depending on the Stress, Histone Deacetylase Inhibitors Act as Heat Shock Protein Co-Inducers in Motor Neurons and Potentiate Arimoclomol, Exerting Neuroprotection through Multiple Mechanisms in ALS Models.” Cell Stress & Chaperones, vol. 25, no. 1, Jan. 2020, pp. 173–91. PubMed, https://doi.org/10.1007/s12192-019-01064-1.

Le Blanc, Katarina, et al. “Treatment of Severe Acute Graft-versus-Host Disease with Third Party Haploidentical Mesenchymal Stem Cells.” The Lancet, vol. 363, no. 9419, May 2004, pp. 1439–41. PubMed, https://doi.org/10.1016/S0140-6736(04)16104-7.

Lee, Dong-Hun, et al. “Neuronal Cell Differentiation of iPSCs for the Clinical Treatment of Neurological Diseases.” Biomedicines, vol. 12, no. 6, June 2024, p. 1350. PubMed, https://doi.org/10.3390/biomedicines12061350.

Li, Shenglan, et al. “Macrophage Migration Inhibitory Factor Suppresses Natural Killer Cell Response and Promotes Hypoimmunogenic Stem Cell Engraftment Following Spinal Cord Injury.” Biology, vol. 14, no. 7, June 2025, p. 791. PubMed, https://doi.org/10.3390/biology14070791.

Li, Xiangyu, et al. “Neural Stem/Progenitor Cell-Derived Extracellular Vesicles: A Novel Therapy for Neurological Diseases and Beyond.” MedComm, vol. 4, no. 1, Feb. 2023, p. e214. PubMed, https://doi.org/10.1002/mco2.214.

Liang, Yi, et al. “The Propensity for Tumorigenesis in Human Induced Pluripotent Stem Cells Is Related with Genomic Instability.” Chinese Journal of Cancer, vol. 32, no. 4, Apr. 2013, pp. 205–12. PubMed, https://doi.org/10.5732/cjc.012.10065.

Liou, Roxanne Hsiang-Chi, et al. “Neuronal Reprogramming for Tissue Repair and Neuroregeneration.” International Journal of Molecular Sciences, vol. 21, no. 12, June 2020, p. 4273. PubMed, https://doi.org/10.3390/ijms21124273.

Lobsiger, Christian S., et al. “Schwann Cells Expressing Dismutase Active Mutant SOD1 Unexpectedly Slow Disease Progression in ALS Mice.” Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 11, Mar. 2009, pp. 4465–70. PubMed, https://doi.org/10.1073/pnas.0813339106.

López-González, Rodrigo, et al. “Transient Recovery in a Rat Model of Familial Amyotrophic Lateral Sclerosis after Transplantation of Motor Neurons Derived from Mouse Embryonic Stem Cells.” Cell Transplantation, vol. 18, no. 10, 2009, pp. 1171–81. PubMed, https://doi.org/10.3727/096368909X12483162197123.

Lunn, J. Simon, et al. “Concise Review: Stem Cell Therapies for Amyotrophic Lateral Sclerosis: Recent Advances and Prospects for the Future.” Stem Cells [Dayton, Ohio], vol. 32, no. 5, May 2014, pp. 1099–1109. PubMed, https://doi.org/10.1002/stem.1628.

Ma, Xiangyu, et al. “Identification of Allogeneic and Xenogeneic Neural Stem Cells’ Immunogenicity in the Brain and Strategies to Alleviate Transplantation Rejection.” Transplant Immunology, vol. 92, Sept. 2025, p. 102247. PubMed, https://doi.org/10.1016/j.trim.2025.102247.

Margeta, Marta. “Neuromuscular Disease: 2023 Update.” Free Neuropathology, vol. 4, Jan. 2023, p. 2. PubMed, https://doi.org/10.17879/freeneuropathology-2023-4682.

Martin, G. R. “Isolation of a Pluripotent Cell Line from Early Mouse Embryos Cultured in Medium Conditioned by Teratocarcinoma Stem Cells.” Proceedings of the National Academy of Sciences of the United States of America, vol. 78, no. 12, Dec. 1981, pp. 7634–38. PubMed, https://doi.org/10.1073/pnas.78.12.7634.

Master, Zubin, et al. “Benefits, Risks and Ethical Considerations in Translation of Stem Cell Research to Clinical Applications in Parkinson’s Disease.” Journal of Medical Ethics, vol. 33, no. 3, Mar. 2007, pp. 169–73. PubMed, https://doi.org/10.1136/jme.2005.013169.

Mattei, Vincenzo, et al. “Validated Methods for Isolation and Qualification of Mesenchymal Stromal/Stem Cells from Different Sources.” Journal of Translational Medicine, vol. 23, no. 1, Sept. 2025, p. 975. PubMed, https://doi.org/10.1186/s12967-025-06972-8.

Mehta, Paul, et al. “Amyotrophic Lateral Sclerosis Estimated Prevalence Cases from 2022 to 2030, Data from the National ALS Registry.” Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration, vol. 26, nos. 3–4, May 2025, pp. 290–95. PubMed, https://doi.org/10.1080/21678421.2024.2447919.

Meyerowitz, Jodi, et al. “C-Jun N-Terminal Kinase Controls TDP-43 Accumulation in Stress Granules Induced by Oxidative Stress.” Molecular Neurodegeneration, vol. 6, Aug. 2011, p. 57. PubMed, https://doi.org/10.1186/1750-1326-6-57.

Mirotsou, Maria, et al. “Paracrine Mechanisms of Stem Cell Reparative and Regenerative Actions in the Heart.” Journal of Molecular and Cellular Cardiology, vol. 50, no. 2, Feb. 2011, pp. 280–89. PubMed, https://doi.org/10.1016/j.yjmcc.2010.08.005.

Mohamadnejad, Mehdi, et al. “Phase 1 Trial of Autologous Bone Marrow Mesenchymal Stem Cell Transplantation in Patients with Decompensated Liver Cirrhosis.” Archives of Iranian Medicine, vol. 10, no. 4, Oct. 2007, pp. 459–66. PubMed, https://doi.org/10.1016/j.bbmt.2007.12.115

Molyneaux, Bradley J., et al. “Neuronal Subtype Specification in the Cerebral Cortex.” Nature Reviews. Neuroscience, vol. 8, no. 6, June 2007, pp. 427–37. PubMed, https://doi.org/10.1038/nrn2151.

Moradi, Sharif, et al. “Research and Therapy with Induced Pluripotent Stem Cells (iPSCs): Social, Legal, and Ethical Considerations.” Stem Cell Research & Therapy, vol. 10, no. 1, Nov. 2019, p. 341. PubMed, https://doi.org/10.1186/s13287-019-1455-y.

Nagamura-Inoue, Tokiko, and Haiping He. “Umbilical Cord-Derived Mesenchymal Stem Cells: Their Advantages and Potential Clinical Utility.” World Journal of Stem Cells, vol. 6, no. 2, Apr. 2014, pp. 195–202. PubMed, https://doi.org/10.4252/wjsc.v6.i2.195.

Nagaya, Noritoshi, et al. “Transplantation of Mesenchymal Stem Cells Improves Cardiac Function in a Rat Model of Dilated Cardiomyopathy.” Circulation, vol. 112, no. 8, Aug. 2005, pp. 1128–35. PubMed, https://doi.org/10.1161/CIRCULATIONAHA.104.500447.

Narayanan, Ramesh K., et al. “Identification of RNA Bound to the TDP-43 Ribonucleoprotein Complex in the Adult Mouse Brain.” Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration, vol. 14, no. 4, May 2013, pp. 252–60. PubMed, https://doi.org/10.3109/21678421.2012.734520.

Nizzardo, Monica, Monica Bucchia, et al. “iPSC-Derived LewisX+CXCR4+β1-Integrin+ Neural Stem Cells Improve the Amyotrophic Lateral Sclerosis Phenotype by Preserving Motor Neurons and Muscle Innervation in Human and Rodent Models.” Human Molecular Genetics, vol. 25, no. 15, Aug. 2016, pp. 3152–63. PubMed, https://doi.org/10.1093/hmg/ddw163.

Nizzardo, Monica, Chiara Simone, et al. “Minimally Invasive Transplantation of iPSC-Derived ALDHhiSSCloVLA4+ Neural Stem Cells Effectively Improves the Phenotype of an Amyotrophic Lateral Sclerosis Model.” Human Molecular Genetics, vol. 23, no. 2, Jan. 2014, pp. 342–54. PubMed, https://doi.org/10.1093/hmg/ddt425.

Okano, Hideyuki, et al. “Ropinirole, a New ALS Drug Candidate Developed Using iPSCs.” Trends in Pharmacological Sciences, vol. 41, no. 2, Feb. 2020, pp. 99–109. PubMed, https://doi.org/10.1016/j.tips.2019.12.002.

Paez-Colasante, Ximena, et al. “Amyotrophic Lateral Sclerosis: Mechanisms and Therapeutics in the Epigenomic Era.” Nature Reviews. Neurology, vol. 11, no. 5, May 2015, pp. 266–79. PubMed, https://doi.org/10.1038/nrneurol.2015.57.

Papadeas, Sophia T., and Nicholas J. Maragakis. “Advances in Stem Cell Research for Amyotrophic Lateral Sclerosis.” Current Opinion in Biotechnology, vol. 20, no. 5, Oct. 2009, pp. 545–51. PubMed, https://doi.org/10.1016/j.copbio.2009.09.003.

Park, Jongjin, et al. “Monoclonal Antibody K312-Based Depletion of Pluripotent Cells from Differentiated Stem Cell Progeny Prevents Teratoma Formation.” BMB Reports, vol. 55, no. 3, Mar. 2022, pp. 142–47. PubMed, https://doi.org/10.5483/BMBRep.2022.55.3.090.

Polymenidou, Magdalini, and Don W. Cleveland. “Prion-like Spread of Protein Aggregates in Neurodegeneration.” The Journal of Experimental Medicine, vol. 209, no. 5, May 2012, pp. 889–93. PubMed, https://doi.org/10.1084/jem.20120741.

Quinn, Colin, and Lauren Elman. “Amyotrophic Lateral Sclerosis and Other Motor Neuron Diseases.” Continuum [Minneapolis, Minn.], vol. 26, no. 5, Oct. 2020, pp. 1323–47. PubMed, https://doi.org/10.1212/CON.0000000000000911.

Ramos-Zúñiga, Rodrigo, et al. “Ethical Implications in the Use of Embryonic and Adult Neural Stem Cells.” Stem Cells International, vol. 2012, 2012, p. 470949. PubMed, https://doi.org/10.1155/2012/470949.

Renton, Alan E., et al. “A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD.” Neuron, vol. 72, no. 2, Oct. 2011, pp. 257–68. PubMed, https://doi.org/10.1016/j.neuron.2011.09.010.

Ringdén, Olle, et al. “Mesenchymal Stem Cells for Treatment of Therapy-Resistant Graft-versus-Host Disease.” Transplantation, vol. 81, no. 10, May 2006, pp. 1390–97. PubMed, https://doi.org/10.1097/01.tp.0000214462.63943.14.

Rippon, H. J., and A. E. Bishop. “Embryonic Stem Cells.” Cell Proliferation, vol. 37, no. 1, Feb. 2004, pp. 23–34. PubMed, https://doi.org/10.1111/j.1365-2184.2004.00298.x.

Ruiz, Sergio, et al. “Limiting Replication Stress during Somatic Cell Reprogramming Reduces Genomic Instability in Induced Pluripotent Stem Cells.” Nature Communications, vol. 6, Aug. 2015, p. 8036. PubMed, https://doi.org/10.1038/ncomms9036.

Sarsenova, Madina, et al. “Mesenchymal Stem Cell-Based Therapy for Rheumatoid Arthritis.” International Journal of Molecular Sciences, vol. 22, no. 21, Oct. 2021, p. 11592. PubMed, https://doi.org/10.3390/ijms222111592.

Schriner, Jacob B., et al. “Low Molecular Weight Heparin Decreases Pro-Coagulant Activity in Clinical MSC Products.” Cytotherapy, vol. 26, no. 2, Feb. 2024, pp. 194–200. PubMed, https://doi.org/10.1016/j.jcyt.2023.11.010.

Shi, Yuzhu, et al. “Engineered Mesenchymal Stem/Stromal Cells against Cancer.” Cell Death & Disease, vol. 16, no. 1, Feb. 2025, p. 113. PubMed, https://doi.org/10.1038/s41419-025-07443-0.

Shigematsu, K., et al. “Long-Term Survival of a Patient with Amyotrophic Lateral Sclerosis (ALS) Who Received Autologous Adipose-Derived Mesenchymal Stem Cells.” European Review for Medical and Pharmacological Sciences, vol. 25, no. 11, June 2021, pp. 4086–90. PubMed, https://doi.org/10.26355/eurrev_202106_26050.

Shoemaker, Lorelei D., and Harley I. Kornblum. “Neural Stem Cells (NSCs) and Proteomics.” Molecular & Cellular Proteomics: MCP, vol. 15, no. 2, Feb. 2016, pp. 344–54. PubMed, https://doi.org/10.1074/mcp.O115.052704.

Singh, Vimal K., et al. “Induced Pluripotent Stem Cells: Applications in Regenerative Medicine, Disease Modeling, and Drug Discovery.” Frontiers in Cell and Developmental Biology, vol. 3, 2015, p. 2. PubMed, https://doi.org/10.3389/fcell.2015.00002.

Siwek, Tomasz, et al. “Repeat Administration of Bone Marrow-Derived Mesenchymal Stem Cells for Treatment of Amyotrophic Lateral Sclerosis.” Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, vol. 26, Dec. 2020, p. e927484. PubMed, https://doi.org/10.12659/MSM.927484.

Srivastava, Amit K., et al. “Serial in Vivo Imaging of Transplanted Allogeneic Neural Stem Cell Survival in a Mouse Model of Amyotrophic Lateral Sclerosis.” Experimental Neurology, vol. 289, Mar. 2017, pp. 96–102. PubMed, https://doi.org/10.1016/j.expneurol.2016.12.011.

Sun, Lingyun, et al. “Umbilical Cord Mesenchymal Stem Cell Transplantation in Severe and Refractory Systemic Lupus Erythematosus.” Arthritis and Rheumatism, vol. 62, no. 8, Aug. 2010, pp. 2467–75. PubMed, https://doi.org/10.1002/art.27548.

Tang, Jingshu, et al. “Umbilical Cord Mesenchymal Stem Cell-Conditioned Medium Inhibits Microglial Activation to Ameliorate Neuroinflammation in Amyotrophic Lateral Sclerosis Mice and Cell Models.” Brain Research Bulletin, vol. 202, Oct. 2023, p. 110760. PubMed, https://doi.org/10.1016/j.brainresbull.2023.110760.

Tang, Yuewen, et al. “Current Progress in the Derivation and Therapeutic Application of Neural Stem Cells.” Cell Death & Disease, vol. 8, no. 10, Oct. 2017, p. e3108. PubMed, https://doi.org/10.1038/cddis.2017.504.

Thomas, K. R., and M. R. Capecchi. “Site-Directed Mutagenesis by Gene Targeting in Mouse Embryo-Derived Stem Cells.” Cell, vol. 51, no. 3, Nov. 1987, pp. 503–12. PubMed, https://doi.org/10.1016/0092-8674(87)90646-5.

Tsuburaya, Naomi, et al. “A Small-Molecule Inhibitor of SOD1-Derlin-1 Interaction Ameliorates Pathology in an ALS Mouse Model.” Nature Communications, vol. 9, no. 1, July 2018, p. 2668. PubMed, https://doi.org/10.1038/s41467-018-05127-2.

Turner, Bradley J., et al. “Dismutase-Competent SOD1 Mutant Accumulation in Myelinating Schwann Cells Is Not Detrimental to Normal or Transgenic ALS Model Mice.” Human Molecular Genetics, vol. 19, no. 5, Mar. 2010, pp. 815–24. PubMed, https://doi.org/10.1093/hmg/ddp550.

Tzeplaeff, Laura, et al. “Current State and Future Directions in the Therapy of ALS.” Cells, vol. 12, no. 11, May 2023, p. 1523. PubMed, https://doi.org/10.3390/cells12111523.

Volarevic, Vladislav, et al. “Ethical and Safety Issues of Stem Cell-Based Therapy.” International Journal of Medical Sciences, vol. 15, no. 1, 2018, pp. 36–45. PubMed, https://doi.org/10.7150/ijms.21666.

Wainger, Brian J., et al. “Intrinsic Membrane Hyperexcitability of Amyotrophic Lateral Sclerosis Patient-Derived Motor Neurons.” Cell Reports, vol. 7, no. 1, Apr. 2014, pp. 1–11. PubMed, https://doi.org/10.1016/j.celrep.2014.03.019.

Wang, Hui, et al. “Recent Progress of the Genetics of Amyotrophic Lateral Sclerosis and Challenges of Gene Therapy.” Frontiers in Neuroscience, vol. 17, 2023, p. 1170996. PubMed, https://doi.org/10.3389/fnins.2023.1170996.

Wang, Lixia, et al. “CRISPR/Cas9-Mediated Targeted Gene Correction in Amyotrophic Lateral Sclerosis Patient iPSCs.” Protein & Cell, vol. 8, no. 5, May 2017, pp. 365–78. PubMed, https://doi.org/10.1007/s13238-017-0397-3.

Wang, Xinxin, et al. “The Pathogenic Mechanism of TAR DNA-Binding Protein 43 (TDP-43) in Amyotrophic Lateral Sclerosis.” Neural Regeneration Research, vol. 19, no. 4, Apr. 2024, pp. 800–06. PubMed, https://doi.org/10.4103/1673-5374.382233.

Wu, Kang-Hsi, et al. “Effective Treatment of Severe Steroid-Resistant Acute Graft-versus-Host Disease with Umbilical Cord-Derived Mesenchymal Stem Cells.” Transplantation, vol. 91, no. 12, June 2011, pp. 1412–16. PubMed, https://doi.org/10.1097/TP.0b013e31821aba18.

Wu, Z., et al. “Thromboembolism Induced by Umbilical Cord Mesenchymal Stem Cell Infusion: A Report of Two Cases and Literature Review.” Transplantation Proceedings, vol. 49, no. 7, Sept. 2017, pp. 1656–58. PubMed, https://doi.org/10.1016/j.transproceed.2017.03.078.

Yang, Yin M., et al. “A Small Molecule Screen in Stem-Cell-Derived Motor Neurons Identifies a Kinase Inhibitor as a Candidate Therapeutic for ALS.” Cell Stem Cell, vol. 12, no. 6, June 2013, pp. 713–26. PubMed, https://doi.org/10.1016/j.stem.2013.04.003.

Yoshihara, Masahito, et al. “Genomic Instability of iPSCs: Challenges Towards Their Clinical Applications.” Stem Cell Reviews and Reports, vol. 13, no. 1, Feb. 2017, pp. 7–16. PubMed, https://doi.org/10.1007/s12015-016-9680-6.

Zakrzewski, Wojciech, et al. “Stem Cells: Past, Present, and Future.” Stem Cell Research & Therapy, vol. 10, no. 1, Feb. 2019, p. 68. PubMed, https://doi.org/10.1186/s13287-019-1165-5.

Zalfa, Cristina, et al. “Transplantation of Clinical-Grade Human Neural Stem Cells Reduces Neuroinflammation, Prolongs Survival and Delays Disease Progression in the SOD1 Rats.” Cell Death & Disease, vol. 10, no. 5, Apr. 2019, p. 345. PubMed, https://doi.org/10.1038/s41419-019-1582-5.

Zeinaloo, Aliakbar, et al. “Intracoronary Administration of Autologous Mesenchymal Stem Cells in a Critically Ill Patient with Dilated Cardiomyopathy.” Pediatric Transplantation, vol. 15, no. 8, Dec. 2011, pp. E183-186. PubMed, https://doi.org/10.1111/j.1399-3046.2010.01366.x.

Zeng, Jixiang, et al. “Decoding TDP-43: The Molecular Chameleon of Neurodegenerative Diseases.” Acta Neuropathologica Communications, vol. 12, no. 1, Dec. 2024, p. 205. PubMed, https://doi.org/10.1186/s40478-024-01914-9.

Zhang, Kehua, et al. “Comprehensive Assessment of the Genomic Stability of Human Induced Pluripotent Stem Cells for Clinical Applications.” Stem Cell Research & Therapy, vol. 17, no. 1, Mar. 2026, p. 171. PubMed, https://doi.org/10.1186/s13287-026-04975-w.

Zhang, Shaoheng, et al. “Comparison of Various Kinds of Bone Marrow Stem Cells for the Repair of Infarcted Myocardium: Single Clonally Purified Non-Hematopoietic Mesenchymal Stem Cells Serve as a Superior Source.” Journal of Cellular Biochemistry, vol. 99, no. 4, Nov. 2006, pp. 1132–47. PubMed, https://doi.org/10.1002/jcb.20949.

Zhao, Xinyu, and Darcie L. Moore. “Neural Stem Cells: Developmental Mechanisms and Disease Modeling.” Cell and Tissue Research, vol. 371, no. 1, Jan. 2018, pp. 1–6. PubMed, https://doi.org/10.1007/s00441-017-2738-1.

Zholudeva, Lyandysha V., et al. “Preparation of Neural Stem Cells and Progenitors: Neuronal Production and Grafting Applications.” Methods in Molecular Biology [Clifton, N.J.], vol. 2311, 2021, pp. 73–108. PubMed, https://doi.org/10.1007/978-1-0716-1437-2_7.

Luxury Logos and Social Identity: How Designer Brands Shape Status Perception

By Mogoș Maria-Alexandra

Abstract

Luxury brands have become powerful symbols of social status, identity, and personal expression, extending far beyond their functional purpose. This paper explores how luxury logos and designer brands influence status perception and contribute to the construction of social identity. Through a literature-based research approach, the study examines the historical development of luxury branding, the psychological motivations behind luxury consumption, and the ways in which luxury products shape both self-perception and public perception.

The research draws on studies from marketing, psychology, and consumer behavior to analyze concepts such as artificial scarcity, social identity theory, self-congruity, and status signaling. A case study of Hermès demonstrates how exclusivity, craftsmanship, and scarcity marketing have enabled the brand to maintain its prestigious position while strengthening consumer desire and the symbolic value of its products. The paper also examines the psychological factors that drive luxury consumption, including the pursuit of status, uniqueness, self-esteem, and social recognition.

The findings suggest that luxury brands function as important social signals that communicate wealth, success, and group affiliation. While luxury consumption can enhance confidence and support identity formation, it may also reinforce materialistic values and social inequality. Overall, the study concludes that luxury branding plays a significant role in shaping both individual identity and societal perceptions of status, making it an influential element of contemporary consumer culture.

Introduction

Despite the exhaustive market overreach of luxury brands and their proportional consumption, a concrete and mutually agreed upon definition of a “luxury brand” has yet to be established (1). The classification of a luxury brand has many underpinnings depending on societal, economic, and political frameworks. Nevertheless, the working assumption is that luxury brands have a strong positive correlation with high quality, prices, exclusivity, and a pervasive image on the market.

A luxury good extends beyond its label of expensive. It is generally perceived by others as special or prestigious. In many cases, the incentive for acquisition is not need-based, but rather it refers to the subliminal message it conveys about social status, annual income, and access to exclusivity.

Luxury brands often rely on artificial scarcity to enhance their perceived status and incentivize purchases. This means that companies purposely limit how many products they sell, even if they could produce more (1). In other words, the product is not naturally rare, but it is intentionally designed to look rare.

Artificial scarcity has a strong effect on how people see luxury products. When something is more difficult to access, people often perceive it as more valuable and desirable. Studies show that when a product's stock is more limited, its price can increase proportionally with desire of consumption, as it feels unique (1). However, if a product is too rare, fewer people can actually buy it, which can reduce how often it appears in public.

Because of this, many luxury brands use artificial scarcity as a strategy. By limiting access to their products, they make them feel more special, which can influence how others see both the product and the person who owns it.

Brands do not receive the luxury classification simply because the cost of their goods and services are high. They reach this status over time, through the manner in which they are marketed and perceived. Research shows that luxury brands focus a lot on image and exclusivity, instead of trying to be universally accessible (2). This is what makes them feel exceptional and different from casual wear.

Overall, luxury brands are not only defined by their high prices or quality, but also by the meanings people attach to them. Through strategies such as artificial scarcity, these brands are able to shape how their products are seen and valued. As a result, luxury goods become more than simple objects, as they start to play a role in how people present themselves and are perceived by others. This connection between luxury consumption and social perception will be explored further in this paper in order to better understand how designer brands influence status and identity.

Methodology

This study employed a literature review methodology to identify, select, and analyse academic sources that explore the relationship between luxury branding and social identity. The purpose of this review is to analyse the connection between luxury logos and status perception, as well as to investigate how designer brands influence the way individuals present themselves and are perceived by others in society.The findings support the primary goal of the thesis: demonstrating that luxury branding, through elements such as logos, exclusivity, and image, plays a significant role in shaping social identity and perceptions of status.

To identify the most relevant and up-to-date evidence, searches were conducted using databases such as Google Scholar and other academic journals. The research period began March 2026 and extended to July 2026.

The keywords used during the process, as well as their various combinations, included "fashion"; “luxury”; “scarcity”; “designer”; “brand”. Boolean operators (AND/OR) were used to narrow or expand search results as the need for more thorough search parameters arose.
In the context of this thesis, the studies spanned between 2014 and 2026 to ensure a comprehensive analysis of the topic and provide credible insights into how luxury brands influence social perception over time. Moreover, only articles published in English were considered as it offered a broader library of resources to formulate compelling findings.

The initial search yielded a library of approximately 80 articles. After a thorough review of their titles and abstracts, only 30 articles were selected for in-depth reviews, of which 12 met the date, methodology, and accuracy criteria established. The following variables were decisive in the process of data extraction from the final pool of studies: type of study, publication year, sample size, focus on luxury branding, analysis of consumer perception, and relevance to social identity and status signaling.

The luxury branding in society literature reviewed revealed trends consistent with the findings of this thesis.

The methodologies applied in these studies demonstrated that luxury logos and designer brands act as important social signals, influencing how individuals are perceived and how they position themselves within social hierarchies.

Historical Overview of Luxury Branding

The concept of luxury can be drawn back to the 14th century, when clothing began to be used to distinguish between social classes (3). During the Middle Ages, laws regulated what people could wear in order to maintain a clear discrepancy between social statuses. Luxury was not merely a means of comfort or pleasure, but a decided identifier of status and reinforcer of differences among social groups.(3)

Over time, the meaning of luxury changed significantly. At first, it was often seen in a negative way, associated with excess, indulgence and aristocracy (3). However, as trade expanded and new goods such as silk, sugar, and spices became available, luxury became more appealing and visible in everyday life (3).

By the 17th and 18th centuries, luxury consumption increased, and specialized shops began to appear, offering high-quality and customized products (3). Later, with the development of transportation and trade, more countries gained access to luxury goods, and local production increased through imitation and adaptation (3).

In the modern period, luxury evolved even further due to industrialization, globalization, and mass media. Luxury brands became more visible, and their image was increased by celebrities, advertising, and global markets (3).

Hermes is one of the earliest examples of a high-end brand, founded in 1837 in Paris. It originally focused on producing harnesses and equipment for the wealthy elite involved in horse riding (3). Over time, the brand expanded into leather goods and clothing, while maintaining a strong focus on quality rather than quantity. This emphasis on scarcity helped Hermes build long-lasting exclusivity and expertise.

Louis Vuitton, founded in 1854, began by producing bespoke travel trunks that were both practical and stylish (3). One of the most important movements in its development was the creation of the LV monogram in 1896, which made the brand easily recognizable (3). Over time, the company extended its product range and became a global symbol of luxury.

Chanel, founded by Coco Chanel, played an important role in changing fashion and luxury branding. She introduced new materials, such as jersey, and created iconic products like Chanel No. 5 perfume and the Chanel classic flap bag (3). The brand's image was strongly influenced by its connection to Hollywood and celebrities, which increased its popularity and visibility.

Burberry is a British luxury brand known for its classic and recognizable designs, especially the trench coat and the iconic plaid pattern (3). For a long time, the brand was associated with a more traditional and conservative image. However, by the late 1980s, this image started to feel outdated, and the brand began to lose its appeal among consumers (3). In response, Burberry changed its strategy by redesigning its products to attract a younger audience while still keeping its original identity (3). It introduced new variations of its signature plaid and expanded its product lines. The brand also adopted digital retailing, which helped it grow and become more relevant in the modern luxury market (3).

Ralph Lauren represents a different approach to luxury branding, focusing on lifestyle rather than just individual products (3). The brand started in 1967 with men's ties, but quickly expanded into a full fashion line (3). One of its most recognizable symbols is the polo player logo, which became strongly associated with a certain lifestyle and image (3).

Over time, Ralph Lauren built its identity around an idealized vision of elegance and success, inspired by movies and popular culture (3). The brand expanded beyond clothing into home products, creating a complete lifestyle concept. This helped it appeal to a wider audience while still maintaining its position in the luxury market (3).

How luxury branding affects self-perception and identity

Luxury branding plays an important role in shaping how individuals see themselves and how they want to be perceived by others. Brand identity is defined as the way a brand communicates its values, personality, and uniqueness, creating a psychological connection between the consumer and the brand (4). This connection goes beyond simple product use, as it allows consumers to feel a sense of belonging and identification with the brand (4). As a result, luxury brands become tools through which individuals express both their internal self, how they see themselves, and their external self, how they want others to see them, (4).

From a social perspective, luxury brands strongly influence self-perception because they are associated with status, wealth, and social class. According to social identity theory, consumers choose products that reflect and reinforce their social identity (4). This means that people are more likely to buy luxury items that associate with the image they want to project in society (4). By observing others, especially reference groups, individuals learn which brands are associated with certain social positions and then imitate those choices to fit in or stand out (4). Therefore, visible elements such as logos become important symbols that signal social status and help individuals create their self-image (4).

In addition to social identity, luxury branding also affects personal identity through the concept of self-congruity. Consumers tend to prefer brands that match their own personality,

values, and lifestyle (4). When there is a strong match between the consumer's self-image and the brand's identity, individuals are more likely to develop emotional attachment and loyalty toward that brand (4). This helps individuals maintain a stable sense of self and increases their confidence in their purchasing decisions (4). In this way, luxury brands are not only external symbols but also internal reinforcements of identity.

Another important factor is the perceived value of luxury brands, which is multidimensional and includes social, emotional, functional, and economic aspects (4). Social value refers to the ability of luxury products to improve one's social image and gain acceptance within different groups (4). Consumers often use luxury goods to impress others and to associate themselves with prestige and exclusivity (4). This directly impacts self-perception, as individuals start to see themselves as part of a higher social class or as more successful and respected (4). Emotional value also plays a key role in shaping identity. Luxury brands create positive feelings such as happiness, confidence, and satisfaction, which influence how individuals perceive themselves (4). These emotional experiences strengthen the bond between the consumer and the brand, making the brand part of the consumer's identity and life (4). For example, when someone feels good wearing a luxury item, they may start associating that feeling with their self-image, reinforcing a more positive perception of themselves (4).

Functional value contributes as well, as consumers associate luxury brands with high quality, durability, and superior design (4). Owning high-quality products can increase self-esteem and create a sense of personal achievement (4). At the same time, these functional attributes support social identity, as high-quality goods are often linked to higher social status (4).

The impact of luxury branding on identity is especially strong during adolescence, a period when individuals are actively forming their sense of self. Teenagers often experience an identity crisis and use consumption as a way to explore and define who they are (5). In this stage, brands play a crucial role, as adolescents choose products that help them fit into desired social groups (5). Owning the "right" brands becomes a way to gain acceptance and avoid social exclusion (5). As a result, luxury brands can significantly influence how teenagers perceive themselves and their place in society (5).

Peer pressure is another important factor that shapes self-perception through luxury branding. Teenagers tend to align their choices with those of their friends, using brands as a shared language to express identity and belonging (5). Branded products are seen as less risky because they are widely recognized and associated with status (5). This makes them a safe option for individuals who are still developing their identity and are concerned about how others perceive them (5). Consequently, luxury branding reinforces both individual and group identity among young consumers (5).

Material possessions, especially luxury fashion items, also contribute to the construction of identity by extending the self (6). Clothing and accessories are not just functional items but also symbolic tools that help individuals define and present themselves (6). Luxury fashion, in particular, allows for self-transformation, enabling individuals to project a more refined or

aspirational version of themselves (6). This shows that identity is not fixed but can be shaped and reshaped through consumption practices (6).

Interestingly, even access to luxury goods without ownership, such as renting or sharing, can influence self-identity. These practices still allow individuals to experience a temporary extension of the self and to present a desired image (6). In some cases, access to luxury items can lead to a more flexible and fluid identity, as individuals can experiment with different styles and social roles (6). This suggests that the symbolic power of luxury branding is not limited to ownership but extends to any form of interaction with the product (6).

In conclusion, luxury branding has a strong impact on both self-perception and identity by influencing how individuals see themselves and how they are perceived by others. Through social, emotional, functional, and economic values, luxury brands shape both personal and social identity. They act as symbols of status, tools for self-expression, and sources of emotional satisfaction. This influence is particularly significant during adolescence but continues throughout life, showing that luxury branding plays a key role in the ongoing construction of identity.

Case Study: The Evolution of Hermes' Luxury Brand Positioning

Hermès is one of the most famous luxury brands in the world today, known for its exclusivity, craftsmanship, and timeless image. The company was founded in Paris in 1837 by Thierry Hermès as a workshop that produced saddles and harnesses for horse carriages (7). During the nineteenth century, horses were one of the main forms of transportation, and Hermès became popular among wealthy customers and the French aristocracy because of the high quality of its handmade products (7). The Hermès logo, which includes a carriage, still represents the company's origins and long history connected to horse riding and craftsmanship (7).

As society and technology changed, Hermès adapted its business instead of remaining only a saddlery company. After the Industrial Revolution, transportation changed and horse carriages became less prevalent, so Hermès started expanding into new luxury products (7). In the 1920s, the company introduced handbags, scarves, belts, and jewelry, and later added perfumes, shoes, clothing, and accessories (7). Over time, Hermès transformed from a small family workshop into an international luxury fashion company while still keeping its focus on traditional handmade production and attention to detail (8).

One of the most important moments in the evolution of Hermès happened in 1956, when actress Grace Kelly was photographed carrying a Hermès handbag in public (7). The photo became very famous, and later the bag was renamed the "Kelly bag" after her (7). This moment helped Hermès become more recognized around the world and connected the brand with elegance, wealth, and social status (7). The Kelly bag quickly became a luxury icon, and owning one became a symbol of prestige and exclusivity (7).

Hermès also built its luxury image through exclusivity and scarcity. The company clearly focuses on wealthy customers instead of trying to appeal to the masses (7). Its high prices automatically limit access to the brand and create the idea that Hermès products are rare and

special(8). Consumers often associate expensive luxury goods with success, prestige, and high social class, which increases the attractiveness of the brand (8). This strategy helped Hermès position itself at the very top of the luxury market.

One of the most famous marketing strategies used by Hermès is scarcity marketing. Popular products like the Birkin and Kelly bags are produced in limited quantities every year, creating a situation where demand is much higher than supply (8). Because these bags are difficult to obtain, consumers often desire them even more (8). Scarcity increases the emotional and symbolic value of the products because owning something rare makes consumers feel unique and important (8). In many social situations, owning a Hermès bag is seen as a sign of wealth, taste, and social recognition (8).

Hermès also uses an allocation system for some of its most exclusive products (8). This means that customers often cannot simply walk into a store and buy a Birkin or Kelly bag immediately (8). In many cases, consumers are encouraged to buy other Hermès products first or build a relationship with the brand before being offered access to limited bags (8). This strategy increases customer loyalty and creates a stronger emotional connection between the consumer and the brand (8). At the same time, it reinforces the image that Hermès products are reserved only for a small and exclusive group of people.

Another reason for Hermès' success is its focus on customer experience and brand culture. The company pays a lot of attention to store design, artistic window displays, and personalized service (7). Hermès stores are carefully designed to reflect the elegance and creativity of the brand (7). The company also organizes exhibitions and collaborates with artists to communicate its values and history through art and storytelling (7). These experiences make consumers feel emotionally connected to the brand and strengthen the luxury image of Hermès. In addition, Hermès has adapted to modern technology and digital communication while still maintaining its traditional image. In recent years, the company expanded into e-commerce and started using digital platforms like WeChat to interact with customers, especially in China (7). The Chinese luxury market became very important for Hermès after the brand opened its first store in Beijing in 1997 (7). As luxury consumption in China increased rapidly, Hermès expanded into more Chinese cities and included Chinese cultural elements in some of its branding and marketing strategies (7). The company also supported charity projects and social responsibility campaigns to improve its public image among Chinese consumers (7).

However, Hermès' scarcity marketing strategy also has some negative effects. Because products are so limited, some consumers buy them only to resell them later at much higher prices (8). This creates long waiting lists and frustration among customers who genuinely want the products (8). Some critics also believe that scarcity marketing creates artificial pressure and encourages impulse buying (8). Despite these criticisms, the strategy continues to be very successful because it maintains the exclusivity and desirability of the brand (8).

The evolution of Hermès' luxury brand positioning shows how the company successfully balanced tradition and innovation over time. Starting as a small saddlery workshop, Hermès became one of the most prestigious luxury brands in the world through craftsmanship,

exclusivity, and strong brand identity(8). The company's focus on handmade quality, scarcity, customer experience, and elite positioning helped create an image of timeless luxury and high social status. Even today, Hermès continues to represent elegance, wealth, and exclusivity, making it one of the strongest luxury brands in the fashion industry.

Hermès' scarcity marketing strategy played a major role in the development of a secondary luxury market. Products like the Birkin and Kelly bags are produced in limited quantities and are very difficult to buy directly from Hermès stores (8). In many cases, customers must wait long periods of time or build a purchase history with the brand before getting access to these bags (8). Because of this, many consumers started searching for alternative ways to buy Hermès products, which led to the rapid growth of the resale market.

The secondhand market for Hermès bags became extremely popular over time because consumers were willing to pay higher prices in order to skip waiting lists and gain immediate access to rare products (9). Auction houses such as Christie's and Heritage Auctions became famous for selling Birkin bags at very high prices, including limited edition and diamond-studded models worth hundreds of thousands of dollars (6). These sales strengthened the idea that Hermès bags are not only fashion accessories, but also collectible items and investments (9).

Online resale platforms also helped expand the secondary market internationally. Websites like The RealReal and BagHunter made it easier for consumers around the world to buy and sell Hermès bags (9). As social media and luxury fashion communities grew, the popularity of resale shopping increased even more. Today, many people buy Hermès bags through resellers instead of directly from boutiques.

Some cities became important hotspots for secondhand luxury shopping. Tokyo is known for its large number of authentic secondhand designer stores and rare Hermès products. Dubai also became a strong luxury resale market because of its wealthy consumers and luxury shopping culture. London is another important center because of its famous consignment stores and auction houses. These cities accelerated the growth of secondary markets entailing the purchase of pre-owned luxury products and made the resale market much more visible internationally. The resale market also increases the cultural status of Hermès bags. They are often viewed as symbols of wealth, prestige, and success instead of simple fashion products (9). Celebrity influence, auction sales, and social media attention continue to increase public fascination with the brand. In this way, the resale market helps keep Hermès highly visible in luxury culture. However, there are also some negative effects. The extremely high resale prices can make some consumers feel excluded or frustrated because Hermès products seem impossible to obtain at normal retail prices (9). Some buyers are also more interested in reselling bags for profit rather than appreciating the craftsmanship or history of the brand. In addition, the popularity of resale markets increases the risk of counterfeit products and reduces Hermès' control over customer experience and pricing.

Overall, Hermès' scarcity marketing strategy directly contributed to the growth of the secondary market. The difficulty of buying Birkin and Kelly bags created strong demand for

resellers, auction houses, and secondhand luxury stores around the world (5)(6). Even though the resale market creates some challenges, it mostly strengthens Hermès' luxury positioning by increasing exclusivity, rarity, and consumer desire for the brand.

A Psychological Analysis of Luxury Consumption

Luxury consumption is about much more than buying expensive products. People often purchase luxury goods because of the emotions, social recognition, and personal satisfaction they provide. According to research, one of the main reasons people buy luxury products is the desire for status and respect from others (10).

Status plays an important role in consumer behavior. Luxury products are often seen as symbols of success, wealth, and achievement. By owning luxury items, consumers can communicate their social position and create a positive impression on others. This is why many people are willing to pay high prices for products that offer prestige in addition to practical value (10).

Another important psychological factor is the desire for uniqueness. Many consumers want to feel different from others and express their individuality through the products they buy. Luxury brands often satisfy this need by offering exclusive products that are rare or difficult to obtain. Owning something that few people can access can make consumers feel special and increase the perceived value of the product (10).

Research also shows that luxury consumption is connected to self-esteem and self-image. Some consumers see luxury purchases as rewards for their hard work and success. Buying luxury products can create feelings of confidence, pride, and personal satisfaction. In this way, luxury goods often have an emotional value that goes beyond their physical function (10).

At the same time, luxury consumption is strongly influenced by society. People do not make purchasing decisions in isolation. Social groups, cultural expectations, and economic conditions all affect consumer behavior. For example, studies have found that in societies with greater income inequality, consumers tend to pay more attention to luxury brands because status differences become more visible and important (10).

Luxury consumption can take different forms. Some consumers prefer products with visible logos that clearly show the brand, while others prefer more discreet products that focus on quality and craftsmanship rather than obvious branding. These differences show that luxury is not only about the product itself but also about the meaning consumers attach to it (10).

Although luxury consumption has several positive effects, it can also have negative consequences. Luxury products may increase confidence and social recognition, but they can also create feelings of guilt or inauthenticity. Some consumers worry that luxury purchases make them appear materialistic or overly focused on status. In addition, people who display luxury goods are sometimes perceived as less approachable or less warm by others (10).

In recent years, the idea of luxury has expanded beyond expensive products. Experiences, education, health, and sustainable lifestyles can also be seen as forms of modern luxury. Many

consumers now seek high status and personal fulfillment through these activities rather than only through material possessions (10).

Overall, luxury consumption is influenced by a combination of psychological and social factors. The need for status, uniqueness, self-expression, and recognition all contribute to consumers' interest in luxury products. Understanding these motivations helps explain why luxury brands continue to attract consumers around the world and why luxury consumption remains an important part of modern society (10).

One important psychological effect of luxury consumption is the perception of competence. People often make assumptions about others based on the products they own and display. When consumers wear luxury brands or use expensive products, observers may associate them with success, intelligence, professional achievement, and financial stability. For example, seeing someone carrying a luxury handbag or wearing an expensive watch may lead others to wonder what profession or business allows them to afford such products. As a result, luxury goods can function as signals of competence and achievement in social and professional environments (10).

Luxury products are also closely connected to status consumption. Research shows that social influence, status-seeking behavior, and materialism are among the strongest factors driving consumers' intentions to purchase luxury goods. Consumers are often motivated by the desire to gain recognition, improve their social image, and communicate success to others. These motivations are especially strong in markets where social status plays an important role in consumer behavior (11).

Another psychological outcome of luxury consumption is pride. Purchasing and owning luxury products can create feelings of accomplishment and personal satisfaction. Consumers may feel proud because luxury items symbolize success, hard work, or achievement. Recent research has shown that pride can positively influence consumers' willingness to purchase luxury products, including sustainable luxury brands. In this case, consumers are not only buying a product but also experiencing positive emotions connected to their identity and self-image (12). These findings demonstrate that luxury consumption is not only about owning expensive products. It also influences how consumers see themselves and how they are perceived by others. Luxury goods can increase feelings of pride, communicate social status, and create an impression of competence, making them powerful psychological symbols in modern society.

Discussion

The findings presented in this paper demonstrate that luxury brands have a significant influence on social identity and status perception. While luxury products are often associated with superior quality and craftsmanship, their value extends far beyond their functional characteristics. Logos, exclusivity, scarcity, and strong brand identities allow luxury brands to function as powerful social symbols that communicate wealth, success, prestige, and belonging. As a result, consumers do not simply purchase luxury goods for practical reasons, but also to express aspects of their identity and shape the way they are perceived by others.

The literature reviewed consistently supports the idea that luxury branding affects both personal and social identity. Concepts such as social identity theory and self-congruity explain why consumers are attracted to brands that reflect the image they wish to project. Luxury brands provide emotional satisfaction, increase self-confidence, and create a sense of belonging to desirable social groups. These effects are particularly evident during adolescence, when identity formation is still developing and peer influence is especially strong.

The case study of Hermès illustrates how luxury brands intentionally create exclusivity through scarcity marketing and controlled distribution. By limiting product availability, Hermès has strengthened the symbolic value of products such as the Birkin and Kelly bags, transforming them into internationally recognized status symbols. The development of a strong secondary resale market further demonstrates that the perceived social value of these products often exceeds their functional value, reinforcing the relationship between scarcity, desirability, and social prestige.

The psychological analysis also shows that luxury consumption is driven by multiple motivations, including the desire for status, uniqueness, competence, and pride. Consumers often use luxury goods as non-verbal signals that communicate success and achievement to others. However, the research also suggests that luxury consumption may produce negative outcomes, such as materialistic values, social exclusion, pressure to maintain a certain image, or judgments based primarily on visible possessions. These findings indicate that luxury branding can strengthen self-esteem for some individuals while simultaneously contributing to social comparison and inequality.

Several limitations should be acknowledged. This paper is based primarily on existing literature rather than original empirical research. In addition, most of the studies reviewed focus on Western luxury markets, while consumer attitudes toward luxury may differ across cultures, age groups, and socioeconomic backgrounds. Future research could include surveys or interviews with consumers from different countries to better understand how luxury logos influence identity and status perception in various cultural contexts. It would also be valuable to examine the growing influence of social media, influencers, and digital luxury marketing on consumer behavior.

In conclusion, luxury brands have evolved into powerful cultural and psychological symbols that shape both individual identity and social perception. Their influence extends beyond fashion and consumption, affecting how people define themselves, interact with others, and communicate social status. As luxury branding continues to evolve in an increasingly digital and globalized world, understanding its impact on identity and society will remain an important area of research in consumer psychology, marketing, and sociology.

Works Cited

1.Peng Xie, Artificial scarcity design heuristics for digital collecting: Evidence from the non-fungible token market, 2025

2.Eunju Ko , John P. Costello , Charles R. Taylor ; What is a luxury brand? A new definition and review of the literature , 21 July 2026

3.Laurie Apple, Leigh Southward, Marianne Bickle; luxury Throughout history: An evaluation of the Industry; 2018

4.Xi Xi, Jing Yang, Kaiwen Jiao, Shanshan Wang, Tianxiang Lu; “We buy what we wanna be”: Understanding the effect of brand identity driven by consumer perceived value in the luxury sector; 15 September 2022

5.Luciana A. Gil, Kyoung-Nan Kwon, Linda K. Good, Lester W. Johnson; Impact of self on attitudes toward luxury brands among teens, 1 March 2011

6.Leila Loussaief, Isabelle Ulrich, Coralie Damay; How does access to luxury fashion challenge self-identity? Exploring women's practices of joint and non-ownership; 14 May 2018

7.Qinnuo X; Proceedings of the 2020 2nd International Conference on Economic Management and Cultural Industry (ICEMCI 2020); 23 October 2020

8.Shuai Wang, The Impact of Scarcity Marketing Models on Consumer Behavior in the Luxury Industry - The Case of Hermes; 2023

9.Brittanny Newsom; Birkin Demand: A Sage & Stylish Investment; 2016

10.David Dubois, Sungjin Jung, Nailya Ordabayeva; The psychology of luxury consumption; 22 July 2020

11.Megha Bharti; The role of socio-psychological and personality antecedents in luxury consumption: a meta-analytic review; 14 December 2021

12.Tahir Islam, Vikas Arya, Ali Ahmed Bodla, Rosa Palladino, Armando Papa; Unmasking luxury consumption and its psychology: An experimental approach to understanding the motivations behind ethical and sustainable brand preferences; 4 July 2024

Evaluating Early-Career BMI Trajectory for Predicting NBA Career Longevity Using Machine Learning By Darsh Goel

Abstract

Nowadays, there are many factors affecting NBA career longevity, and one of them is body composition, which is often mentioned but not often investigated through the early seasons of a player's career. This research is focused on finding whether BMI trajectory, which refers to the change and variation of a player's Body Mass Index across his first three recorded seasons, may serve as a predictor of the remaining career length. The *NBA Players* dataset (Cirtautas) is suitable for this type of investigation since it allows longitudinal tracking of each player's development and therefore helps build their trajectory over multiple seasons. The final modeling sample included 1,197 players. To conduct the analysis, the final players were used for constructing BMI trajectory features and other early-career variables such as draft information, demographic variables, and performance variables. The hypothesis was that BMI trajectory during a player's first three recorded seasons would improve prediction of remaining career length beyond standard early-career variables. According to the results, the hypothesis was not supported: the model using standard draft, demographic, and performance variables achieved an R^2 of 0.454. Adding the first-season BMI produced the same R^2, while adding the BMI trajectory features gave an R^2 of 0.448. These results suggest that the BMI trajectory features used in this study did not provide clear additional predictive information beyond standard early-career variables.

I. Introduction

NBA career longevity refers to the total length of time a player remains in professional basketball. Since the NBA started in 1946, the league has formed into one of the most competitive and professional sports in the world (NBA.com Staff). The players face constant physical, athletic, and mental pressure. Throughout NBA history, the number of seasons a player plays varies a lot: some play more than a decade, and some play for a couple of years before they retire. These differences have made career longevity an important topic because it reflects durability, adaptation, and long-term development. The NBA has changed across eras in terms of pace, physicality, training, and style of play. This makes predicting career outcomes more difficult.

Predicting an NBA player's career length is one of the more complex and studied challenges in sports analytics. Career length is affected by many factors, including player performance, draft position, age, injuries, team opportunity, and physical development, which makes prediction harder. Recently, the use of Machine Learning (ML) models has shown promising progress in addressing this complexity because they can reveal non-linear relationships and learn from large amounts of historical data (Pedregosa et al.).

The traditional statistical models, while helpful, often fail to represent how player development changes over time. In contrast, machine learning models such as Random Forest,

Gradient Boosting, and other methods have become useful tools for predictive sports analysis. Many existing works related to the NBA focus on performance-based variables or draft information, while very few works focus on the body composition of a player and how it changes through the early stages of a career. Body Mass Index (BMI) is usually considered at one point in time rather than a trajectory across multiple seasons. Because of this, the BMI trajectory may provide additional information than the static variables.

This study applies multiple machine learning models to investigate whether BMI trajectory during a player's first three seasons provides useful information for predicting remaining NBA career length. The hypothesis of this study is that the BMI trajectory during a player's first three seasons will improve the prediction of the remaining NBA career length beyond standard early-career variables. To test this, prediction performance was analyzed using several models, including a Linear Regression baseline and more advanced machine learning models. This study also examines whether the BMI trajectory provides additional predictive information when combined with draft, demographic, and performance data.

II. Prior Research and Research Gap

Previous research related to the NBA and career outcomes has focused on variables such as draft position, age, scoring ability, and other performance-based statistics. According to Moxley and Towne, age, the college that the player went to, and college performance were useful in predicting early NBA career success. Their research also found that physical and athletic measurements helped predict the draft order. However, these measurements were less effective when trying to separate productive NBA players from unproductive players after their previous performance was evaluated.

Injuries are one of the biggest factors in the NBA and career longevity. Kester et al. studied NBA players who had their anterior cruciate ligament (ACL) reconstructed and found that these players had shorter careers and lower performance after returning compared to similar players who never went through that injury. Their findings show that physical health can affect performance and how long a player remains in professional basketball. However, studies that are based on injuries focus on one specific time instead of analysing the physical change that happens to a player.

Body composition has also been studied in basketball players. Fields et al. examined seasonal and long-term changes in the body composition of NCAA Division 1 basketball athletes. Their study showed that a player's body composition could change based on playing position and across seasons and academic years. Strength and conditioning programs are suggested by these researchers to help respond to these changes. However, this study focuses on college athletes and did not investigate whether body composition changes could be used to predict career longevity.

Researchers have been using machine learning and data-mining methods to study NBA performance. These methods can analyze demographic, performance, economic, and injury-related variables all at the same time. Research by Sarlis and Tjortjis used sports analytics

to examine how age, position, injuries, and performance are connected. This research shows that machine learning can help identify patterns and relationships between many variables, which may be difficult with simpler analysis.

The existing research gives useful information about early-career performance, draft position, and body composition. However, these areas are often studied differently. Most researchers use body composition as a single point in time, or for college athletes, while limited research uses it as a trajectory to predict future NBA career length. This leaves a gap that this study focuses on: whether a player's BMI changes through his early NBA seasons may provide additional information beyond standard early-career variables.

III. Data and Methodology

A. Dataset and Player Selection

The dataset used in this study was the *NBA Players* dataset created by Justinas Cirtautas and published on Kaggle (Cirtautas). The dataset includes NBA player data from 1996 to 2022. Each row represented a player from one NBA season and included information such as height, weight, age, games played, points, rebounds, assists, net rating, draft round, and draft number. These variables represented important information that could be measured during the early seasons of a player's career.

Each player appeared in multiple rows, which allowed the model to track the players across different seasons. The records were grouped by the player's name and then arranged in chronological order. This made it easy for the model to figure out each player's first three recorded NBA seasons and calculate the change in performance and body composition over that period.

Players with at least three recorded NBA seasons were included in the final analysis. Other players were removed because they did not have enough data to create a three-season BMI trajectory. After this process, the final dataset included 1,197 players. Since the dataset recorded players from 2022 and some of those players are still active, they were also removed from the final dataset because their complete career lengths were still unknown.

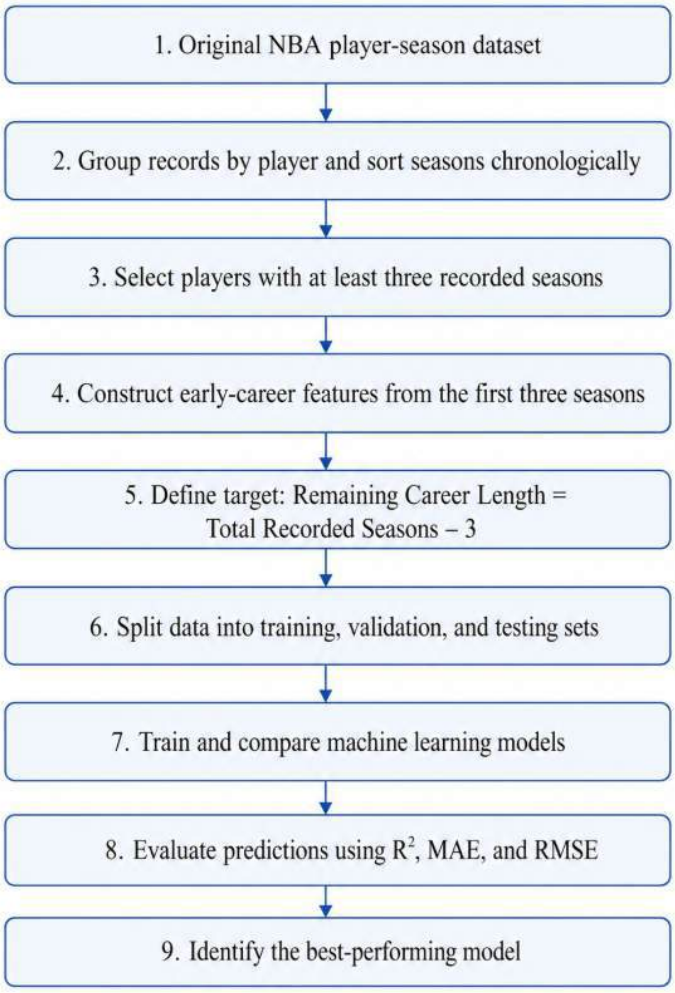


Figure 1. Overview of the research process.

B. Constructing Early-Career Features

The BMI of each player was calculated using each player's listed height and weight. The player height was converted from centimeters to meters before using the standard BMI formula:

$$BMI = \frac{Weight\ in\ Kilograms}{(Height\ in\ Meters)^2}$$

(National Heart, Lung, and Blood Institute)

In this study, the BMI trajectory is the change and variation of a player's BMI within his first three seasons. It was represented by using his first-season BMI, third-season BMI, BMI change, and BMI standard deviation. BMI change was calculated as the difference between the third-season start and the first-season BMI:

$$BMI\ Change = 3rd\ Season\ BMI - 1st\ Season\ BMI$$

A positive BMI change represented that the player's BMI has increased across his first three seasons, while a negative value showed that it decreased. However, BMI change only measured the overall difference between the periods and did not indicate how stable or unstable his BMI was during the seasons in between.

BMI standard deviation is the measured amount of variation in a player's BMI across his first three seasons and was represented as bmi_std. A higher value indicated that the player's BMI had greater variation, while a lower value showed that his BMI stayed stable. BMI change and BMI standard deviation both allowed the model to analyze the direction of the BMI development and the amount of variation that occurred during the early seasons.

In addition to the BMI trajectory features, standard career variables were created. Draft variables included draft round, draft number, and whether a player went undrafted. Demographic variables included the player's age and height at the beginning of his NBA career. These variables helped because they may represent the player's early expectations and development.

Performance features were also created using each player's first three seasons. The variables included were games played, points, rebounds, assists, and net rating. The average was taken for each variable and used to calculate a player's overall early-career production. Changes for all statistics were also included to show whether a player's performance improved or declined.

To determine if the BMI trajectory added value to the prediction, three main feature groups were created. The first feature group consisted of the draft, demographic, and performance variables; the second group contained the same standard variables with the first-season BMI; and the third model included the standard variables and the BMI trajectory features. The same modeling and evaluation procedures were used for all three groups so that their prediction performances could be compared fairly.

C. Building Predictive Models

In this study, five predictive models were trained and compared: Linear regression, Random Forest Regression, Extra Trees Regression, Gradient Boosting Regression, and Histogram-Based Gradient Boosting. The models were implemented in Python using the scikit-learn machine-learning library (Pedregosa et al.). These models were selected because they allowed the study to compare traditional models with more complex machine learning models that can analyze non-linear relationships.

The Linear Regression model was used as the traditional comparison model. This model estimates the relationship between the feature variables and the target variable by using a linear combination. This model's interpretability is simpler than that of the other models. However, Linear Regression's accuracy when figuring out relationships between non-linear variables or multiple variables could be less effective.

The Random Forest Regression model builds a group of Decision Trees using different samples of the training data and different combinations of features. The final result comes from the average of the predictions from these trees. Since this model uses many Decision Trees, it

reduces the instability that happens when using only one Decision Tree. This also allows the model to identify more complex relationships. However, this model can still be inaccurate because it could overfit when trees become too complex or when the dataset contains noise.

Extra Trees Regression is very similar to the Random Forest Regression model since they both combine predictions from Decision Trees. Extra Trees adds additional randomness when selecting the data to divide, which could reduce model variance and improve generalization. However, this randomness could cause the model to miss a few valuable patterns when the divisions are too arbitrary.

The Gradient Boosting Regression model constructs Decision Trees one at a time. Each new Decision Tree learns from the prediction errors made by the previous trees. This allows the model to gradually improve, and it can also capture the relationship between BMI trajectory, draft information, demographic characteristics, and performance statistics. If the model uses too many trees, it can overfit and provide results that aren't accurate.

Histogram-Based Gradient Boosting Regression follows a similar boosting process, but histogram-based refers to the model grouping continuous values into intervals. This can make the training faster and more efficient, while still analyzing the non-linear relationship and patterns. This method is more useful when the dataset contains many numerical features and a large number of observations. Together, these ensemble models use multiple decision trees or boosting processes to identify complex relationships in numerical data ("Ensembles").

Before the model was trained, the remaining career length was transformed using a logarithmic transformation:

$$y_{transformed} = log(1 + y)$$

This transformation was done to reduce the effect of players with unusually long remaining careers and make the target distribution less uneven. After each model gave its predictions, the values were transformed back to the original scale:

$$\hat{y} = e^{\hat{y}transformed} - 1$$

Any negative predictions were changed to zero, since the remaining career length cannot be a negative number.

D. Measuring Model Performance

To evaluate the model performance clearly, the dataset was divided into training, validation, and test sets using an 80/10/10 split. Each model was trained using the same training, validation, and testing sets so that their predictions could be compared fairly. The training set was used to fit each model, and the validation set was used to compare the settings and to select the strongest model configuration. The testing set was kept separate until the final analysis so the model performance could be measured on unseen player data.

Three versions of the model dataset were used to determine whether the BMI trajectory added predictive value. The same model type, data split, and evaluation process were used for all three versions. This allowed the standard-variable model to be compared with the BMI trajectory features model fairly.

Three evaluation metrics were used in this study: R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics were used to evaluate the explained variation and prediction error on unseen player data ("Metrics and Scoring").

R-squared measured the amount of variation in the remaining career length that was explained by the model ("Metrics and Scoring"). It was calculated using this formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)2}{\sum_{i=1}^{n} (y_i - \bar{y})2}$$

In this formula, y_i represents the actual remaining career length, $\hat{y}_i$ represents the predicted remaining career length, and $\bar{y}$ represents the average actual remaining career length. An R^2 value that is closer to 1 shows that the model explained more of the variation between players. A value closer to 0 indicated that the model predicted similarly to the average remaining career length.

Mean Absolute Error measured the average absolute difference between the predicted and the actual number of remaining seasons ("Metrics and Scoring"):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

A lower MAE indicated that the model's performance was closer to the actual remaining seasons left. Since the MAE was measured in seasons, it gave a clear way to understand the average size of the model's prediction errors.

Root Mean Squared Error also measured the difference between predicted and actual career lengths, but it gave more weight to the larger prediction errors ("Metrics and Scoring"):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)2}$$

A lower RMSE value showed that the model performance was stronger. Since the large errors were squared before averaging, RMSE helped identify models that gave predictions that were far from the actual number of remaining seasons.

All predictions were converted back to the original number of seasons before R^2, MAE, and RMSE were calculated. This made sure that the calculated errors represented actual NBA seasons and could be interpreted more clearly. A stronger model would show a higher R^2 value and lower MAE and RMSE values.

IV. Results

A. Comparing Initial Model Performance

The first experiment compared the performance of all five models using the same training and validation datasets. Each model was trained using the combined feature group, which included all the standard variables and the BMI trajectory features. The models were first trained using the training set and then evaluated on the validation set using R^2, MAE, and RMSE.

Linear Regression produced the strongest initial validation performance. The model achieved an R^2 of 0.552, which meant that it explained 55.2% of the variation in the remaining career length. Its MAE was 1.823 seasons, and its RMSE was 2.446 seasons. Although Linear Regression was the simplest model used in the study, it performed much better than the complex tree-based models. This showed that the relationship between the standard variables and BMI trajectory features could be displayed using a linear model.

The Extra Trees Regression model performed similarly to the Linear Regression model with an R^2 of 0.546, an MAE of 1.856 seasons, and an RMSE of 2.462 seasons. This model explained slightly less variation in the remaining career length compared to the Linear model, but it showed that a tree-based model could also identify patterns in the early-career data.

Random Forest Regression achieved an R^2 of 0.527, an MAE of 1.895, and an RMSE of 2.514 seasons. Although its performance was weaker than that of the Linear Regression and Extra Trees models, it still explained more than half of the variation in the remaining career length.

Gradient Boosting Regression and the Histogram-Based Gradient Boosting models produced weaker results than the rest. Gradient Boosting achieved an R^2 of 0.495, an MAE of 1.961 seasons, and an RMSE of 2.596 seasons. Histogram-Based Gradient Boosting achieved an R^2 of 0.484, an MAE of 1.930 seasons, and an RMSE of 2.626 seasons. These models were able to find the non-linear patterns, but their initial settings did not generalize to unseen data as well as Linear Regression, Extra Trees Regression, and Random Forest Regression.

A baseline model, which predicted the average remaining career length, was also evaluated. The baseline model achieved an R^2 of -0.026, an MAE of 3.178 seasons, and an RMSE of 3.703 seasons. All five models performed better than the baseline model, which shows that the standard early-career variables were useful information for predicting the remaining NBA career length.

Table 1: Initial Validation-Set Performance Comparison

Model	R^2	MAE	RMSE
Linear Regression	0.552	1.823	2.446
Extra Trees Regression	0.546	1.856	2.462
Random Forest Regression	0.527	1.895	2.514

Gradient Boosting Regression	0.495	1.961	2.596
Histogram-Based Gradient Boosting	0.484	1.930	2.626
Baseline Model	-0.026	3.178	3.703

B. Optimized Gradient Boosting & Effect of Adding BMI Trajectory

Although Linear Regression produced the strongest initial validation performance, the Gradient Boosting Regression model was selected for further optimization because its initial settings had not been adjusted to match the dataset. This model can identify non-linear relationships between BMI trajectory, draft information, demographic characteristics, and performance statistics. Different combinations of the model settings were tested, such as the number of trees, learning rate, tree depth, and the minimum number of samples needed to divide a node. The final settings were determined by the validation performance and not the test set to avoid overfitting.

The optimized Gradient Boosting model was used to test whether BMI information provided more predictive value. Three different versions of the model were compared. The first model used only standard draft, demographic, and performance variables. The second model used the same variables with the first-season BMI added. The third model used the standard variables with the BMI trajectory feature group. All three models used the same model settings, data split, and evaluation process so that their performances could be compared fairly.

The model using only standard variables achieved an R^2 of 0.454. Adding first-season BMI produced the same R^2 of 0.454, which showed that the BMI at one point left the R^2 unchanged and produced only negligible changes in the MAE and the RMSE. The third model, when the complete BMI feature group was added, achieved an R^2 of 0.448. The difference between the BMI trajectory model and the standard variable model was:

$$0.448 - 0.454 = -0.006$$

These results did not support the hypothesis of the study. The standard-variable model performed slightly better than the BMI trajectory model. Therefore, the model using standard early-career variables performed the best among all three Gradient Boosting feature configurations.

Table 2: Effect of Adding BMI Trajectory Features

Feature Group	R^2	MAE	RMSE
Standard Early-Career Variables	0.454	2.197	2.895
Standard Variables + First-Season BMI	0.454	2.193	2.894
Standard Variables + BMI Trajectory	0.448	2.199	2.911

C. Which Features Mattered Most

Permutation feature importance was used to examine which variables mattered the most in the optimized Gradient Boosting model. This measured how much the model's R^2 decreased when the values of one feature were randomly rearranged. A larger decrease showed that the model depended more strongly on that feature, while a value closer to zero showed that the feature had little to no effect on the prediction performance ("Permutation Feature Importance").

The most important feature was average games played during the first three recorded seasons, or gp_avg, with an average importance value of 0.1847. This meant that if this feature were to be randomly arranged, it would cause the largest decrease in the model's test R^2. The next important features were starting age, average points, and change in games played, with importance values of 0.1283, 0.0551, and 0.0465, respectively. These results suggest that the model depended most strongly on information related to opportunity, age, and early performance.

For the BMI trajectory features, BMI standard deviation had the highest importance value of 0.0133. The other BMI variables, including first-season BMI, third-season BMI, and BMI change, had importance values of 0.0013, -0.0011, and -0.0108, respectively. Only the BMI standard deviation appeared in the ten most important variables. However, its importance value was much smaller compared to the rest of the leading variables. This finding is consistent with the feature-group comparison because adding the BMI trajectory features decreased the model's held-out test performance.

The importance of each variable result also showed that the model relied on many different types of early-career information rather than one single variable alone. Draft, demographic, and performance variables each contributed to the predictions, but their levels of contribution were different.

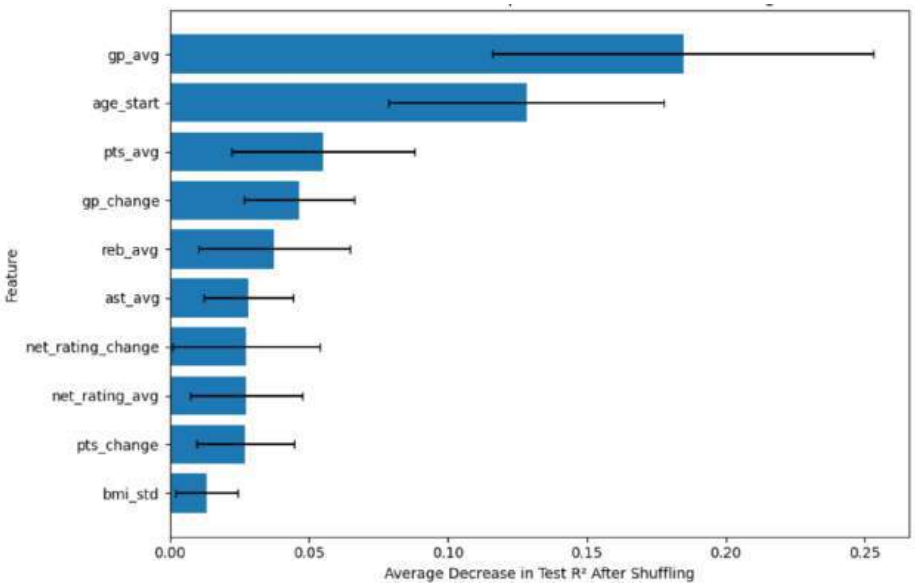


Figure 2. Permutation importance of the ten most important variables

V. Discussion

The main finding of this study was that the BMI trajectory does not improve the prediction of remaining NBA career length. Since the difference was small, the results do not show that BMI trajectory reduced prediction performance, but they showed that the BMI trajectory features did not provide additional information beyond the standard variables.

One possible explanation for this is that the standard variables already represented many of the factors connected to career longevity. Draft position could reflect how a team sees a player's potential before entering the league. The performance statistics may describe the player's early opportunities. These variables already contain useful information about ability, role, and team expectations. The BMI trajectory features may not have added enough new information to improve the model.

Another possible reason is that BMI does not fully show physical development. A BMI increase could result from muscle growth, increased body fat, injury recovery, or changes in training. A decrease could result from improved conditioning, a change in position, illness, or loss of muscle mass. Because of this, two players with the same BMI change could have experienced different physical development.

The Gradient Boosting Regression model helped test whether BMI trajectory could be useful through non-linear patterns. This model could identify whether a BMI change had different effects on a player depending on age, draft position, starting BMI, or early performance. However, even with such a flexible model, the BMI trajectory features did not improve the predictions on the held-out test set.

The findings in this study also connect with previous research on NBA development and career outcomes, as already mentioned. Moxley and Towne showed that performance, age, and college information could help predict NBA success. Kester et al. demonstrated that physical health and major injuries could affect performance and career lengths. Fields et al. found that the body composition could change across seasons in college basketball players. However, the results of the current study suggest that BMI alone may be too general to represent the physical factors described in the previous research.

From a practical perspective, the study shows that collecting information over multiple seasons does not make it a useful predictor. The team and analysts could gain more useful information from body-fat percentage, lean muscle mass, strength, conditioning, and injury history than from BMI trajectory alone. Therefore, future research should use more detailed variables on physical development when studying the relationship between body development and NBA career longevity.

VI. Limitations and Future Directions

While this study gave useful findings, several limitations need to be mentioned. First, BMI does not provide a full measurement of the body composition of a player. BMI is calculated with the height and the weight of a player. BMI does not separate muscle mass from body fat (National Heart, Lung, and Blood Institute). This is very important for NBA players because

many players have high body weights due to muscle mass rather than fat. Two players with similar BMI values could have very different physical builds.

Another limitation is that the dataset did not include other factors that affect career longevity, such as injuries, playing position, team role, contract decisions, medical history, and personal decisions. All of these variables can affect how long a player stays in the NBA. For example, a player might have to leave the league early due to a severe injury even though his early-career stats and BMI values appear strong.

The study also included only players with at least three recorded seasons. This was important because the model needed three seasons to calculate BMI change and BMI standard deviation. However, this means that the model does not apply to players whose careers ended before their third recorded season. The players who were in the dataset with fewer than three seasons and are still playing are also not included.

Several improvements could also make this model stronger for future research. A dataset with more physical variables, like body-fat percentage, lean muscle mass, wingspan, strength measurements, and conditioning results, could give a more accurate description of a player's body composition. Other variables like injury history, position, minutes played, games started, contract status, and team role could also be added to explain how long a player stays in the league.

Future studies could test other modeling techniques and longer early-career periods. For example, researchers could compare BMI development across two, three, four, or five seasons and determine which period produced the strongest results. More advanced models could predict if a player stays in the league after each season, rather than predicting one final career length.

VII. Conclusion

This study examined whether the BMI trajectory during a player's first three recorded NBA seasons could improve the prediction of a player's remaining career length. Five different predictive models were tested on the same dataset using draft, demographic, performance, and BMI trajectory features. The Gradient Boosting Regression model was then optimized and used to compare the predictive value of different feature groups.

The model using the standard draft, demographic, and performance variables achieved a held-out test R^2 of 0.454. Adding the first-season BMI produced the same R^2 and only small changes in MAE and RMSE, and adding the BMI trajectory features decreased the R^2 to 0.448. The findings did not support the original hypothesis.

More detailed variables such as body-fat percentage, lean muscle mass, injury reports, and playing position could provide a better understanding of how body developments affect career longevity. Overall, the study showed that the early draft, demographic, and performance information was useful for predicting career longevity, while BMI trajectory did not provide additional predictive value for the final model.

Works Cited

Cirtautas, Justinas. “NBA Players.” Kaggle, https://www.kaggle.com/datasets/justinas/nba-players-data. Accessed 16 July 2026.

“Ensembles: Gradient Boosting, Random Forests, Bagging, Voting, Stacking.” Scikit-learn, https://scikit-learn.org/stable/modules/ensemble. Accessed 16 July 2026.

Fields, Jennifer B., et al. “Seasonal and Longitudinal Changes in Body Composition by Sport-Position in NCAA Division I Basketball Athletes.” Sports, vol. 6, no. 3, 2018, article 85, https://doi.org/10.3390/sports6030085.

Kester, Benjamin S., et al. “Athletic Performance and Career Longevity Following Anterior Cruciate Ligament Reconstruction in the National Basketball Association.” Knee Surgery, Sports Traumatology, Arthroscopy, vol. 25, no. 10, 2017, pp. 3031–37, https://doi.org/10.1007/s00167-016-4060-y.

“Metrics and Scoring: Quantifying the Quality of Predictions.” Scikit-learn, https://scikit-learn.org/stable/modules/model_evaluation. Accessed 16 July 2026.

Moxley, Jerad H., and Tyler J. Towne. “Predicting Success in the National Basketball Association: Stability & Potential.” Psychology of Sport and Exercise, vol. 16, part 1, Jan. 2015, pp. 128–36, https://doi.org/10.1016/j.psychsport.2014.07.003.

“Calculate Your BMI.” National Heart, Lung, and Blood Institute, U.S. Department of Health and Human Services, https://www.nhlbi.nih.gov/calculate-your-bmi. Accessed 16 July 2026.

NBA.com Staff. “NBA Frequently Asked Questions.” NBA.com, 20 Mar. 2019, https://www.nba.com/news/faq. Accessed 16 July 2026.

Pedregosa, Fabian, et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–30, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

“Permutation Feature Importance.” Scikit-learn, https://scikit-learn.org/stable/modules/permutation_importance.html. Accessed 16 July 2026.

Sarlis, Vangelis, and Christos Tjortjis. “Sports Analytics: Data Mining to Uncover NBA Player Position, Age, and Injury Impact on Performance and Economics.” Information, vol. 15, no. 4, 2024, article 242, https://www.mdpi.com/2078-2489/15/4/242.

From the Course to the Simulator: Off-Course Golf and the Technology-enabled Experience EconomyBy Le Tuan Minh

Abstract

This paper examines how the growth of off-course golf is transforming golf as both a sport and a commercial industry. Off-course golf includes simulators, technology-enabled driving ranges, golf entertainment venues, and virtual golf courses. Drawing on academic literature and selected industry reports, this review reveals the three major drivers of the growth of the off-course segment: the use of new technologies, gamification, and an increased demand for social and entertainment-led experiences. This new approach to golf reduces difficulties connected with time, weather, and financial aspects of the game, allowing casual participants and consumers with limited golf experience to participate. This trend is encouraging hybrid business models that combine equipment sales with the recurring revenue from memberships, software applications, subscriptions, and data-dependent services. The article also underlines a number of new concerns regarding ownership and commercialization of performance data, consumer privacy, and dependence on closed digital platforms. These trends suggest that the new opportunities arising from increasing engagement carry commercial risks as well. In summary, golf is expanding the community of participants and transforming the customer experience, sales, and interaction with this sport. Therefore, golf is moving towards a hybrid business approach combining athletes' performance with the aspects of technologies, entertainment, and service consumption but, at the same time, raises a number of challenges related to sustainability and data management as well as questions regarding the deep relationships of off-course and on-course modes of golf play.

Introduction

Historically, golf has been understood as a course-centered sport linked with country club memberships and time-intensive rounds on outdoor golf courses. The traditional model of golf participation is known as "on-course golf," characterized by playing golf on a physical golf course, which requires substantial commitments of time, money, and skill development. In fact, for many years, this model defined the structure and culture of the golf industry, where consumer demand was centered on clubs, golf apparel, golf balls, membership fees, and other on-course golf-related services. More recently, the golf industry is experiencing a transformation brought by advancements in technology and the emergence of off-course golf. Off-course golf can be defined as any form of golf activity carried out away from traditional golf courses, including indoor golf simulators, entertainment venues such as Topgolf, virtual golf experiences that use launch monitors, and simulation technology (National Golf Foundation [NGF], 2025).

The growth of off-course golf reflects a broader trend. Traditional golf has frequently been met with issues associated with high costs, lack of accessibility, land needs, and significant time investment. On the contrary, off-course golf is characterized by greater convenience and accessibility, enabling consumers to enjoy the sport in urban centers, indoors, and over relatively

shorter periods of time. Innovations in golf such as simulation software, artificial intelligence for swing coaching, and launch monitoring technologies are just some of the ways golf is transforming from a primarily course-based sport into a hybrid sport that combines sports participation, entertainment, and technology.

Recent statistics demonstrate the significant growth of this transformation. According to the National Golf Foundation (NGF), the total U.S. golf participation reached an all-time peak of 48.1 million participants in 2025 when combining both on-course and off-course exclusive participants and dual participants. The NGF estimates that participation could touch 50 million by the end of 2026 if it grows by a minimum of 4% (National Golf Foundation, 2026). Off-course-golf-only participants alone reached 19.1 million Americans in 2024, which includes participation at entertainment venues, indoor simulators, and tech-enabled driving ranges. The overall U.S. golf participants increased by 41% over the past six years, but the number of on-course-golfers-only fell to 12.1 million players in 2023 from 13.2 million in 2022, displaying a decline of more than a million participants. By the end of 2025, that number continued to decrease, as the number of on-course-only participants dropped to 10.2 million golfers (NGF, 2026; Retailstat Financial, 2025). During the same period, the population of dual participants, meaning those who engage in golf either on or off the course, rose from 12.4 million in 2022 to 14.5 million in 2023 and 19 million by 2025. This observation implies that the decrease in on-course participation does not correspond to golfers leaving the sport. Instead, these trends suggest that off-course golf is not trying to replace traditional golf but is expanding the industry by creating more accessible entry points into the sport and increasing engagement with golf culture and consumption.

The expansion of off-course golf also has important implications for the composition of consumer demand. Conventional golf has historically been structured around course membership and physical equipment such as apparel, clubs, or balls. Off-course formats generate demand for a wider ecosystem of technology-enabled products and services, such as simulators, launch monitors, and AI swing aids. Accordingly, the industry is evolving in two connecting ways: how technology is altering how golf is played and experienced, while consumer spending is shifting toward digital tools, recurring services, and entertainment-based participation.

The paper examines how technology-enabled off-course golf is transforming participation, consumer demand and business models within the golf industry. It argues that off-course formats are expanding participation while restructuring demand towards digital products, subscription-based recurring services, and experience-based consumption. The paper draws on academic studies of both on-course and virtual golf and sport consumption and industry reports on participation, sustainability, and market development to explore how golf is evolving from a traditional course-centered activity into a more accessible, technology-enabled, and entertainment-oriented industry.

Defining off-course golf

Off-course golf refers to golf-related participation taking place outside conventional golf courses. These include utilizing the indoor golf simulator; entertainment facilities such as

Topgolf; high-tech driving ranges; and virtual golf with the aid of launch monitors and AI-enabled swing analysis techniques (National Golf Foundation, 2025). Off-course golf is considered to be a much different sphere compared to traditional golf due to the use of technology-based solutions and the environment that makes golfing a more group-oriented activity. Nonetheless, off-course golf should not be seen as a homogeneous market because it serves different consumer expectations, business models, and usage contexts.

An important distinction with off-course golf is between entertainment-focused participation and performance-focused participation. Entertainment-focused venues promote social interaction, leisure activities, and accessibility through multiplayer games, food, music, and casual experiences. In contrast, performance-focused formats, such as simulator studios and launch-monitor systems, are primarily used for practice, data analysis, and personalized skill development. These technologies provide consumers with detailed swing metrics and feedback in order to improve performance, creating a performance-driven environment. Consequently, off-course golf does not generate a single type of consumer demand; rather, it creates multiple consumer markets characterized by different motivations, preferences, and spending behaviors.

The emergence of these formats has also changed consumer expectations within the golf industry. In contrast to on-course golf, off-course consumers increasingly value convenience, entertainment, flexibility, and digital engagement (Retailstat Financial, 2025). Therefore, businesses have adapted by investing in simulator technology, AI-supported coaching systems, entertainment venues, and subscription-based services. Existing studies have examined golf entertainment venues primarily as entry points into the sport and as mechanisms for increasing participation. For example, Sears and Cianfrone (2025) argue that golf entertainment venues serve as the "front porch to golf," introducing new consumers to the game and expanding participation beyond traditional golf courses. Research has also examined social presence and behavioural intentions in virtual simulators, technology-based self-service, screen-golf decision-making, technology acceptance and the relationship between alternative and traditional golf (Lee et al., 2013). Nevertheless, relatively little attention has been paid to understanding how these distinct forms of off-course participation are reshaping consumer demand for golf products and services. This gap highlights the need for further research into the implications of off-course golf for both consumer behavior and the future structure of the golf industry.

Why off-course golf has grown

The growth of off-course golf is largely due to three key factors: technology and gamification, accessibility and convenience, and entertainment and social interaction. Unlike conventional golf played on the course, off-course golf is less time-consuming, less expensive, and requires less access to vast outdoor areas. As a result, off-course golf has grown to become a significant hub of innovation in the industry as consumers' preferences shift toward playing golf with the help of technology, convenience, and experience-driven leisure activities.

The first important element of the growth of off-course golf is technology and gamification. Gamification has been framed as the design of information systems that use

game-related features to create gameful experiences and enhance users' motivation and engagement (Koivisto and Hamari, 2019). The emergence of various devices aimed at improving golf performance, like golf simulators, ball-tracking systems, and launch monitors has drastically improved the effectiveness of swing analysis and shot tracing. The technology supplied by golf simulators improves the golf experience for beginners, as players can now practise golf in controlled, indoor environments throughout the year.

Accessibility and convenience are other vital factors influencing the emergence of off-course golf. Traditional round golf involves substantial time commitments paired with large land usage, which restricts the likelihood of people from metropolitan cities participating in it. Even beyond accessibility, many environmental concerns are associated with the land management, water usage, and sustainability challenges faced by golf courses (BIGGA, n.d.). In contrast, off-course golf has minimal space requirements, which means that it is located in closer proximity to the customer. Nonetheless, the off-course alternative cannot be regarded as the ultimate solution for the sustainability issues of golf. It still requires electricity, equipment maintenance, and technology upgrades. According to research (Belkhir & Elmeligi, 2018), information and communication technologies also have an environmental impact in terms of energy consumption, which means that off-course golf may overcome some land access constraints but also create new sustainability issues.

Off-course golf does not only increase physical accessibility; it also provides an improvement in psychological accessibility by eliminating the intimidation that results from conventional golf. For first-time players, visiting an entire golf course for play can be a challenging experience because they have to deal with concerns regarding pace of play as well as complex regulations and requirements surrounding golf equipment conventions and embarrassment that comes about. Off-course formats offer a comfortable and controlled environment where individuals can learn without feeling the pressure that comes from traditional golf culture. This is important because studies conducted on virtual golf indicate that this type of experience reduces barriers not only with respect to time, cost, and weather, but also with respect to skill confidence (Choi et al., 2019). Thus, off-course golf allows for increased participation in golf because it keeps it accessible not only physically but also socially.

The third main driver is social experiences and entertainment. Venues like Topgolf are making golf less formal by emphasizing group play, food, music, and social experience over traditional competition. The method makes golf accessible to consumers with little or no golfing experience. Research examining golf entertainment venues similarly shows that the social environment and emotional experience make golf more attractive to beginners (Sears & Cianfrone, 2025). Consequently, the activity has transcended from the traditional sport practice to a wider technology-based sector.

From Equipment Sales to Platform-Based Golf Consumption

The growth of off-course golf is reshaping patterns of consumer demand within the golf industry. Rather than simply increasing demand for traditional golf equipment, off-course

participation has expanded the market for technology-enabled products and transformed golf into a broader ecosystem of experiences and services. Platform-based golf consumption occurs when hardware, software, performance data, coaching, virtual content and consumer communities are connected through a shared digital system. Consequently, golf consumption is no longer centered solely around physical products but increasingly around digital technologies, subscriptions, analytics, and entertainment-based experiences.

First, off-course golf has expanded demand for technology-enabled products, effectively transforming portions of the golf equipment market into a hardware-software-service market. The rise of simulators, launch monitors, AI-assisted coaching systems, and swing analysis technologies exemplifies how much people want items that help enhance performance by providing immediate feedback and personalized training. According to Choi et al. (2019), virtual golf eliminates a number of obstacles that one faces when playing golf outdoors, such as price, climate, time and level of skill, which means that technological solutions can make golf affordable for more people. Likewise, Breedlove et al. (2023) have found that launch monitor technology positively influences consumers in terms of learning, shopping habits, and expenditures on golf gear. As simulator participation continues to increase, value creation is shifting beyond physical clubs and balls toward integrated digital platforms and analytics systems. Industry analysis also indicates that software and service revenues are becoming increasingly important within the simulator market, suggesting that future competition may depend as much on software ecosystems as on hardware capabilities (Fortune Business Insights, 2026).

Second, off-course golf has created demand for experiences rather than only products. Entertainment venues such as Topgolf and simulator lounges have transformed golf by increasing demand for experiences characterized by simulator bookings, memberships, food and beverage sales, league competitions, and corporate events. According to Sears and Cianfrone (2025), golf entertainment venues serve as accessible entry points into the sport by offering social and leisure experiences that appeal to consumers beyond traditional golfers. The study reveals that novice golfers observed at Topgolf prioritize socialization, favorable sports landscapes, feelings of excitement, and the place serving as a “front porch” onto the golf world. Thus, the golf industry is now more focused on leisure and entertainment ideas and looks for more ways of generating profit through various events and subscriptions instead of equipment sales only.

Third, off-course golf has broadened product demand among beginners and casual consumers by creating new entry points into the sport. Unlike traditional on-course golf, which usually caters to committed players, experience-focused formats attract consumers with different motivations and lower levels of experience. Choi et al. (2019) provide strong empirical evidence that virtual golf reduces major participation barriers, especially time, cost, weather, and skill confidence, making golf more accessible than traditional on-course play. This accessibility does not only increase participation; it also changes what consumers buy. Beginners and casual players are more likely to demand entry-level clubs, rental equipment, affordable training aids,

portable launch monitors, swing-analysis applications, simulator memberships, beginner lesson packages, and group entertainment bookings. Sears and Cianfrone (2025) similarly show that golf entertainment venues can improve novice golfers' perceptions of the sport by making golf feel more social, convenient, and less intimidating. As a result, off-course golf expands demand for products and services designed around accessibility, instruction, convenience, and social participation rather than only advanced athletic performance.

These changes have significant implications for business models within the golf industry because off-course golf connects experiences directly to product demand. Simulator lounges, Topgolf-style venues, and virtual coaching platforms do not only sell entertainment. They expose consumers to golf equipment, training technology, and performance data in a low-pressure setting. As beginners and casual players participate in these experiences, they may become more likely to purchase entry-level clubs, gloves, balls, apparel, lessons, launch monitors, swing-analysis apps, simulator memberships, and other golf-related products. Sears and Cianfrone (2025) show that golf entertainment venues improve beginner-level golfers' perception of the sport, while Choi et al. (2019) demonstrate that virtual golf reduces barriers such as time, cost, weather, and skill confidence. Therefore, off-course golf can act as both an entertainment product and a gateway into broader golf consumption. This means that experience-based participation does not replace product demand. Instead, it reshapes the sport by turning social and digital golf experiences into new demand for equipment, technology, and subscriptions.

This increasing reliance on digital golf platforms creates critical questions about how golf businesses will make money in the future and how dependent consumers may become on recurring access. Alongside purchases of clubs, balls, and one-time lessons, consumers are increasingly paying for access to platforms, simulator software, virtual courses, swing-analysis tools, memberships, and coaching subscriptions. This does not mean that the entire physical golf equipment market will disappear, but, instead, that traditional product sales are increasingly being supplemented by recurring service-based revenue. Although focused on a non-golf context, Nansubuga and Kowalkowski (2024) explain that subscription-based models allow product-centered firms to develop service-based revenue streams alongside existing product sales. In the context of off-course golf, this means companies can generate recurring revenue not only through physical equipment but also through monthly memberships, platform access fees, content upgrades, virtual-course libraries, coaching applications, and data-driven training services. While this model can create more stable revenue for businesses, it may also increase consumer dependence on closed software ecosystems, where continued access to performance tools requires ongoing payment.

As the digitalization of golf is becoming apparent, one of the most important but overlooked issues is the growing value of consumer data. Every swing taken in a simulator or launch-monitor system can produce detailed performance metrics, including ball speed, club speed, launch angle, attack angle, and shot dispersion. As off-course golf keeps growing, simulation companies, training platforms, shops, and makers of different kinds of equipment may start having access to the information regarding consumers' performances. This data will

improve both training and the selection of the clubs that a particular person is supposed to use. At the same time, however, this data will provide companies with opportunities for selling particular products to consumers. Thus, performance data creates value to both consumers and companies who are looking forward to influencing the consumers' future purchases and use of the platform. At the same time, research regarding fitness or self-tracking technologies indicates that consumers create important pieces of information without being aware of how this data would be used, with it being connected to storage, analysis, or monetization of performance data (Alhajri et al., 2022; Lupton, 2021). Thus, with the expansion of off-course golf, it is important to address ethical issues linked to the data belonging to the consumer as well. Evidence on the relationship between off-course and traditional golf is mixed. Alternative-golf satisfaction and involvement may encourage traditional-golf participation, yet highly valued screen-golf experiences may also reduce intentions to play on-course. Off-course golf may therefore function simultaneously as a gateway, a complement, and a substitute.

Conclusion

This review has argued that off-course golf is expanding rather than replacing traditional golf. These days, the main aspect of the golf industry lies in traditional on-course golf, which has golf courses, equipment sales, memberships, and competitive events. Nevertheless, modern consumers do participate in golf differently, with new practices available to them. Among such practices are golf simulators, entertainment venues (for example, Topgolf), launch monitors, AI-powered coaching tools, and golf simulators. For this reason, golf becomes more accessible and social for casual players and first-time participants to the game, especially since it becomes the hybrid business that fuses active participation, technology, and entertainment.
This change is also reshaping consumers' demands for golf. These days, consumers are using subscription services and participating in group events, as well as tools and services like performance analytics and digital coaching. Such services do complement and enhance the consumption demands for traditional golf, such as golf clubs, balls, sports apparel, green fees, and coaching. Now it becomes clear that golf businesses will have to start using new business models that combine physical goods sales with subscriptions, software services, venue memberships and data-driven platforms.

In any case, the long-term implications remain uncertain for the industry in perspective. It remains uncertain whether consumers engaging in off-course activities will choose to participate in physical golf courses any longer or if it remains a leisurely activity. Additionally, there are some questions about who has access to golfers' performance data and whether it is utilized for commercial purposes or not. Furthermore, there is a danger that subscription models may create financial lock-in or dependency on proprietary platforms and digital platforms and will cease to be able to benefit from golf-related products and services in the event they stop paying. In the end, even though off-course golf has led to the increase of engagement in the sport, it has also transformed the process of consuming golf as a product and service.

Works Cited

Alhajri, M., Salehi Shahraki, A., & Rudolph, C. (2022). Privacy of fitness applications and consent management in blockchain. In M. Ngoc Dinh (Ed.), Proceedings of 2022 Australasian Computer Science Week (ACSW 2022) (pp. 65–73). Association for Computing Machinery. https://doi.org/10.1145/3511616.3513100

Belkhir, L., & Elmeligi, A. (2018). Assessing ICT global emissions footprint: Trends to 2040 & recommendations. Journal of Cleaner Production, 177, 448–463. https://doi.org/10.1016/j.jclepro.2017.12.239

Breedlove, J., Chen, S. S., Wilson, E. C., Su, Y., & Zhang, J. J. (2023). Golfer responses to traditional and technology-enabled equipment sales: A comparative study. International Journal of Golf Science, 11(1). https://www.golfsciencejournal.org/article/90593-golfer-responses-to-traditional-and-technology-enabled-equipment-sales-a-comparative-study

British and International Golf Greenkeepers Association. (n.d.). Sustainable golf. https://www.bigga.org.uk/about/green/sustainable-golf.html

Choi, C., Greenwell, T. C., Hums, M. A., & Hambrick, M. E. (2019). Understanding consumer behaviors in virtual golf: Differences in leisure constraints. Sport Marketing Quarterly, 28(1), 46–57. https://doi.org/10.32731/SMQ.281.032019.04

Fortune Business Insights. (2026, June 1). Golf simulator market size, share & industry analysis, by offering, product type, simulator type, business model, end-user, and regional forecast report, 2026–2034. https://www.fortunebusinessinsights.com/golf-simulator-market-110374

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. International Journal of Information Management, 45, 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Lee, H. G., Chung, S., & Lee, W. H. (2013). Presence in virtual golf simulators: The effects of presence on perceived enjoyment, perceived value, and behavioral intention. New Media & Society, 15(6), 930–946.

Lupton, D. (2021). "Sharing is caring": Australian self-trackers' concepts and practices of personal data sharing and privacy. Frontiers in Digital Health, 3, Article 649275. https://doi.org/10.3389/fdgth.2021.649275

Nansubuga, B., & Kowalkowski, C. (2024). Moving to subscriptions: Service growth through business model innovation in consumer and business markets. Journal of Service Management, 35(6), 185–215. https://doi.org/10.1108/JOSM-10-2023-0438

National Golf Foundation. (2026a). Golf industry facts. https://www.ngf.org/the-clubhouse/golf-industry-research/

National Golf Foundation. (2026b, March 6). Golf's growth era—The road to 50 million golfers. https://www.ngf.org/short-game/golfs-growth-era-the-road-to-50-million-golfers/

RetailStat Financial. (2025, March 18). Golf industry 2024 recap [Special analysis]. Creditntell. https://www.creditntell.com/reports/67758.pdf

Sears, J., & Cianfrone, B. A. (2025). Golf entertainment venues as the front porch to golf: Exploring the experiences and perceptions of novice golfers at Topgolf. Sports Innovation Journal, 6, 43–55. https://doi.org/10.18060/29238

The Dark Side of the Pygmalion Effect By Vaishali Sita Noronha

1. Abstract

The Pygmalion effect describes the influence of expectations on individual performance through self-fulfilling prophecies (Eden, 1992). While extensively studied in educational contexts (Rosenthal & Jacobson, 1968; Brophy & Good, 1970; Jussim & Harber, 2005), its application within contemporary work organisations has received comparatively limited critical attention, particularly with regard to potential negative outcomes. As organisational cultures increasingly emphasise performance, employee monitoring, and productivity, workplace conditions may intensify the psychological pressures associated with expectancy-based practices (Ravid et al., 2020). This paper examines the research question: *What are the negative implications of artificially inducing the Pygmalion effect as a management tool in 21st-century work organisations?* A narrative literature review was conducted to synthesise existing research. Relevant studies were identified through searches on Google Scholar using key terms related to the Pygmalion effect with preference given to recent peer-reviewed literature. This review found that, despite its frequent portrayal as a positive managerial strategy, the Pygmalion effect is associated with unintended consequences such as burnout, ethical challenges, biased performance evaluations, and psychological strain within modern workplaces. By critically examining these underexplored dimensions, the paper aims to contribute to a more balanced understanding of expectancy effects and inform more ethically conscious organisational practices.

Keywords Pygmalion Effect, Leadership, Work Organisations, Performance Pressure, Expectancy Effect, Self-fulfilling Prophecy, Expectations, Stress

2. Introduction

The Expectancy Effect is a psychological phenomenon in which an individual's beliefs or expectations about another person or situation unconsciously influence the other person's behaviour, thereby shaping outcomes (Rosenthal & Jacobson, 1968). One well-known manifestation of this phenomenon is the Pygmalion Effect, which describes the relationship between higher expectations and enhanced performance (Rosenthal & Jacobson, 1968). One of the earliest and most influential demonstrations of the Pygmalion effect was conducted by Rosenthal and Jacobson (1968), and the phenomenon has since been widely studied within educational settings and, more recently, applied to organisational contexts, where it is frequently framed as a beneficial tool for improving motivation and performance (Eden, 1992).

The Pygmalion effect, however, has been subject to ongoing debate and criticism (Jussim & Harber, 2005). In light of these concerns, this paper examines the research question: *What are the negative implications of artificially inducing the Pygmalion effect as a management tool in 21st-century work organisations?* Much of the existing literature has prioritised the examination of positive performance outcomes (Eden, 1992), often overlooking three key areas of concern.

First, there is growing concern regarding the psychological consequences of expectancy-based practices, including stress, pressure, and emotional strain, which may be exacerbated by contemporary workplace conditions (Timmermans & Rubie-Davies, 2018). These concerns are particularly prevalent in 21st-century work environments defined by heightened performance demands and blurred boundaries between work and personal life (Bakker & Demerouti, 2007). For younger generations entering the workforce, including Millennials and Generation Z, expectations are often continuously communicated and evaluated, potentially intensifying their psychological impact (Twenge, 2017). Second, the effectiveness of the Pygmalion effect depends heavily on the interpersonal abilities of managers and employees, limiting its practical applicability across organisational settings (Eden, 1992). Finally, the ethical implications of intentionally employing the effect as a management tool have received comparatively less scholarly attention than its performance outcomes (Jussim & Harber, 2005).

Accordingly, this paper aims to examine the negative implications of artificially inducing the Pygmalion effect in modern work organisations. Through a narrative review of the literature, the paper focuses on the potential impact of the Pygmalion effect on employee mental health, its limitations and contextual variability, and the ethical considerations surrounding their application, in order to provide a more critical and balanced understanding of the Pygmalion effect in contemporary organisational settings.

3. Method

This study employed a narrative review methodology to examine the negative consequences of the Pygmalion effect across workplace contexts. A narrative approach was selected to allow for a broad analysis of existing literature spanning leadership, followership, and organisational behaviour. A literature search was conducted using Google Scholar as the primary database. Searches were performed using combinations of the following keywords: “Pygmalion Effect,” “Leadership,” “Followership,” “Workplace Expectations,” “Self-fulfilling Prophecy,” “Burnout” and “Organisational Performance”. Reference lists of relevant articles were also reviewed to identify additional sources not captured in the initial search.

Peer-reviewed journal articles written in English and focusing on workplace or organisational settings were considered. Selected articles were reviewed to identify recurring themes related to the adverse effects of the Pygmalion effect. Findings were synthesised to highlight patterns and gaps in the existing literature.

4. Findings

4.1 Impact of the Pygmalion Effect on Psychological Well-being

While high expectations are often framed in organisational and psychological literature as motivational forces that guide behaviour and enhance performance, recent research suggests that they can also increase psychological strain (Mitchell et al., 2019; Xu et al., 2021). Performance pressure is defined as “the urgency to achieve high-performance levels because performance is

tied to substantial consequences" (Mitchell et al., 2019, p. 533). Expectations can lead to performance pressure, particularly when outcomes are tied to an individual's self-worth (Crocker & Wolfe, 2001).

Xu et al. (2021) conducted a quantitative survey of 345 employees across various Chinese organisations and found that increased performance pressure, associated with the Pygmalion effect, led to greater self-objectification among employees. Self-objectification occurs when individuals begin viewing themselves as instruments for achieving workplace goals rather than as autonomous individuals (Gruenfeld et al., 2008; Baldissarri et al., 2014). Supporting these concerns, Baldissarri et al. (2014) found that workplace objectification was associated with greater emotional exhaustion, reduced psychological well-being, and increased depressive symptoms among employees.

The Pygmalion effect may also create psychological concerns when high managerial expectations generate excessive performance pressure. Because the Pygmalion effect relies on employees internalising and striving to meet elevated expectations, employees may experience anxiety about whether they can achieve the standards set by their leaders. Xu et al. (2021) found that performance pressure is positively associated with workplace anxiety and that workplace anxiety negatively affects employee functioning. This suggests that employees may experience heightened stress, worry, and fear of failure when attempting to meet elevated expectations. Rather than serving solely as a source of motivation, managerial expectations may therefore become a source of psychological strain, particularly when employees perceive the expected standards as difficult to attain. The Pygmalion effect is traditionally understood to arise from managers' genuine beliefs in employees' capabilities (Eden, 1992). However, when used intentionally as a performance-enhancement technique, expectations may become less reflective of employees' actual abilities and more focused on achieving organisational outcomes. As a result, expectations may be more likely to exceed what employees realistically perceive themselves as capable of achieving, intensifying performance pressure and the associated psychological strain identified by Xu et al. (2021).

Ichilov (2025) argues that the successful operation of the Pygmalion effect depends upon employees maintaining a "peace state," characterised by emotional regulation, reduced stress, and cognitive clarity. In the absence of such conditions, heightened expectations may increase psychological strain and impair employees' ability to process and respond constructively to workplace demands (Ichilov, 2025). As employees internalise managerial expectations and incorporate them into their self-perceptions, expectations that are perceived as rigid or continually escalating may intensify performance pressure and anxiety (Rosenthal & Jacobson, 1968; Ichilov, 2025). Deteriorating mental health may in turn reduce an employee's ability to meet organisational expectations, creating a vicious downward spiral, increasing psychological distress, and progressively worsening performance.

A possible explanation for these negative psychological consequences lies in the impact of chronic stress on cognitive functioning. Processing high expectations, learning, and adaptive problem-solving rely on the brain's prefrontal cortex to guide goal-directed behaviour

(Goldman-Rakic, 1996; Miller & Cohen, 2001). However, when a manager imposes high standards on an employee whose nervous system is already under chronic stress, the brain may misinterpret those expectations as an existential threat rather than an inspiring challenge. As Ichilov (2025) notes, a true peace state requires "reduced amygdala activation" to optimise the neural networks required for growth. Without this calm baseline, catecholamines and glucocorticoids (neurochemicals associated with the body's stress response) rapidly weaken prefrontal connections, reducing an individual's capacity for top-down cognitive control (Arnsten, 2009). The employee loses the cognitive clarity needed to process and act upon positive expectations. Instead of motivating the worker, the manager's demands simply flood the nervous system with panic, forcing a state of cognitive paralysis (Arnsten, 2009).

Not only can those subject to the Pygmalion effect suffer negative mental health effects, but also the individuals in positions of power employing this management technique (Timmermans & Rubie-Davies, 2018, p. 395). Supporting this claim, Timmermans and Rubie-Davies (2018) conducted a large-scale quantitative survey across multiple classrooms and found that teachers who differentiated more strongly between students in their expectations reported significantly higher levels of classroom stress and lower psychological need satisfaction. This effect was particularly evident for relatedness, which refers to the fundamental need to feel connected to, understood by, and maintain positive relationships with others (Ryan & Deci, 2017).

The pressure to constantly monitor, categorise, and respond differently to students may transform expectations from motivational tools into stressors. This suggests that expectation-driven performance systems may undermine the well-being of those responsible for implementing them even as they may improve the short-term performance of those subjected to them. Although conducted within an educational setting, these findings may hold relevance for workplace environments, where managers similarly evaluate performance, establish expectations, and monitor the progress of subordinates.

4.2 Limitations in the Effectiveness of the Pygmalion Effect

If the Pygmalion effect is to be considered a viable organisational strategy, an important question must first be addressed: Can expectations be deliberately shaped without any legitimate basis or evidence of capabilities? It is important to note that attempts to raise managerial expectations do not always succeed. When supervisors possess substantial prior knowledge of employee competence, existing perceptions become resistant to change and require especially convincing information about an employee before they can be altered (Eden, 1992). This limitation was illustrated by Sutton and Woodman (1989), who attempted to induce a Pygmalion effect in retail staff. They informed retail sales managers that based on personality test results, certain employees possessed "exceptional sales potential"; however, these“exceptional employees” were randomly selected. This study failed to produce the anticipated results. Eden (1992) suggests that this failure may have occurred because managerial expectations were never meaningfully elevated, preventing the self-fulfilling prophecy mechanism from operating.

Even if it is possible to successfully alter expectations, it does not guarantee the emergence of a Pygmalion effect. Drawing on an integrative review of previous research into interpersonal expectancy effects, Cooper and Hazelrigg (1988) found that expectancy effects were influenced by the characteristics of both the inducer and the target. Managers who possessed a stronger desire to influence others, communicated more effectively through nonverbal cues, and were better liked by employees were more likely to generate expectancy-congruent behaviour. Likewise, employees who were more receptive to social influence and better able to interpret interpersonal signals were more likely to conform to managerial expectations (Cooper & Hazelrigg, 1988). Building upon Eden's thesis, they concluded that expectancy effects are highly contingent upon interpersonal factors rather than occurring automatically.

Beyond interpersonal characteristics, the effectiveness of the Pygmalion effect is also dependent on favourable situational and environmental conditions. Far from being a universal catalyst for achievement, the Pygmalion effect is usually less effective if the job context contradicts a manager's demands, turning a motivational framework into an engine of workplace exhaustion (Chen et al., 2023). According to Chen et al. (2023), a manager's high expectations fail to translate into successful outcomes in two distinct workplace contexts: those that deny operational freedom and those involving unmanageable workloads.

The Pygmalion framework cannot be effectively applied when a manager's high expectations are paired with less operational freedom. According to Chen et al. (2023), when a supervisor sets high expectations but restricts freedom through rigid oversight, it creates a severe "cue inconsistency" that disrupts the employee's psychological processing and paralyzes their capacity for innovation. For example, when a tech manager asks an engineer to build a highly creative new app feature, but then micromanages the process; the rigidity denies the employee "job autonomy," causing them to freeze up and preventing innovation.

The Pygmalion effect also loses effectiveness if the job design buries employees under crushing workloads and tight timelines. When an employee faces intense performance pressure alongside excessive work demands, it creates a direct resource conflict that drains their mental bandwidth and prevents them from translating supervisor expectations into actual "creative effort" (Chen et al., 2023). A clear example of this occurs when an accounting manager asks for flawless, innovative financial audits from a team already working long 14-hour days during peak tax season; because the excessive "work demands" completely drain their cognitive energy, the high expectations fail to inspire and instead cause panic and lessen baseline efficiency.

4.3 Ethics regarding usage of the Pygmalion Effect

The intentional application of the Pygmalion effect within organisational settings raises ethical concerns regarding the way it influences the allocation of opportunities among employees. In a majority of the literature, the Pygmalion effect is defined as a self-fulfilling prophecy whereby high expectations held by an authority figure lead to improved performance by the target of those expectations (Eden, 1992; Rosenthal & Jacobson, 1968). However, much

of the experimental research employs the Pygmalion effect in a way that results in differential expectations between participants. Weinstein (2018) argues that this differentiation in expectations not only shapes individuals' self-views through the internalisation of these expectations but also results in the differential allocation of opportunities. Individuals for whom higher expectations are held receive greater access to support, guidance, and opportunities for development, creating advantages that may subsequently translate into improved performance outcomes, while others receive comparatively fewer resources and developmental opportunities (Weinstein, 2018). Over time, the stronger performance of the favoured group may appear to validate the manager's expectations, when it may instead reflect the unequal allocation of organisational resources. Consequently, the intentional use of the Pygmalion effect may contribute to inequitable talent development, as professional success becomes influenced by the subjective expectations that determine access to organisational opportunities.

Although experimental research demonstrates the benefits of induced high expectations, their application in organisational settings is more complex, as managerial expectations are often influenced by implicit bias and unconscious stereotypes rather than objective assessments of employee potential. As the Pygmalion effect relies on leaders first forming expectations before communicating them to employees, biases present at this initial stage may shape who is perceived as having high potential. Heilman (2001) argues that gender stereotypes influence perceptions of competence and leadership ability, causing equally capable individuals to be evaluated differently because they do not conform to stereotypical expectations of success. Similarly, leaders' implicit conceptions of followers have been shown to influence the expectations they develop about employees, which subsequently shape leader-follower interactions and employee performance (Whiteley, Sy, & Johnson, 2012). Consequently, when organisations intentionally encourage managers to employ expectancy-based management practices, the employees selected to receive high expectations may reflect unconscious stereotypes rather than objective potential. Rather than operating as a merit-based motivational strategy, the intentional application of the Pygmalion effect may therefore reinforce existing organisational inequalities by amplifying biases already present in managerial decision-making.

Importantly, this process is not unidirectional. Employees' implicit biases and stereotypes also shape the perceived credibility of their leaders, influencing the degree to which managerial expectations are internalised. Karakowsky et al. (2017) argue that expectancy effects are influenced by "gender-based contextual cues" that shape how followers perceive a leader's authority and expertise. Consequently, female leaders may be less successful in intentionally employing the Pygmalion effect if employees are less likely to regard their expectations as legitimate because of negative gender stereotypes. This argument is consistent with broader leadership research, which has found that women in positions of authority are often evaluated as less competent or less authoritative than men despite equivalent performance, reducing the influence of their leadership behaviours. If the effectiveness of a management strategy depends, in part, on social stereotypes surrounding the identity of the leader, its application cannot be considered ethically neutral. Rather than functioning as a universally applicable management

tool, the intentional use of the Pygmalion effect may advantage leaders who conform to traditional leadership stereotypes while placing women and other minority and under-represented groups at a systematic disadvantage.

5. Conclusion

This paper sought to examine the research question: *What are the negative implications of artificially inducing the Pygmalion effect as a management tool in 21st-century work organisations?* The evidence reviewed indicates that the intentional application of the Pygmalion effect should not be regarded as a universally effective or ethically neutral management strategy. Although artificially elevating managerial expectations can improve employee performance under certain, favourable conditions, its effectiveness depends on situational, interpersonal, and organisational factors rather than occurring consistently across organisational settings. More importantly, its deliberate application may also generate significant psychological, practical, and ethical consequences.

These findings have important implications for organisational practice. As 21st-century workplaces continue to emphasise productivity, high performance, and continuous evaluation, organisations may be tempted to adopt Pygmalion effect-based management strategies as a means of improving employee outcomes. However, the evidence reviewed suggests that simply encouraging managers to communicate higher expectations is unlikely to consistently produce positive results. Artificially induced expectations may place excessive pressure on employees, particularly when expectations exceed perceived capabilities or are introduced within demanding work environments. Furthermore, the effectiveness of the Pygmalion effect depends heavily on the interpersonal skills of managers and the willingness of employees to internalise managerial expectations, making it difficult to apply consistently across different organisational contexts. Organisations should therefore exercise caution if incorporating Pygmalion effect-based practices into leadership training.

6. Limitations

The conclusions drawn from this review should also be considered in light of several limitations within the existing body of research. Much of the foundational literature on the Pygmalion effect originates from educational settings, with comparatively fewer studies examining its intentional application in contemporary workplaces. It is uncertain whether findings derived from classrooms can be fully generalised to organisational environments, where managerial relationships, organisational structures, and performance demands differ. Furthermore, contemporary organisational research examining the intentional application of the Pygmalion effect is sparse. As a result, it was not possible to limit the literature review to contemporary sources. Instead, older foundational studies often examining different organisational contexts had to be relied upon.

A further limitation of the Pygmalion literature concerns the manner in which expectancy effects have traditionally been induced. In the original study by Rosenthal and Jacobson (1968),

teachers were informed that a randomly selected group of students were likely to experience exceptional intellectual growth, leading them to hold higher expectations for these students than for their classmates. Eden (1992) adopted a similar approach within organisational settings by informing supervisors that selected employees possessed particularly high potential, thereby encouraging managers to communicate elevated expectations towards only a subset of workers. As a result, it becomes difficult to determine whether the reported improvements in performance result from the communication of higher expectations itself or from the differential allocation of managerial attention, encouragement, feedback, and developmental opportunities afforded to those identified as high potential. The observed outcomes may therefore reflect the combined effects of heightened expectations and preferential treatment, making it difficult to isolate the independent contribution of expectations alone. This distinction has important implications for organisations seeking to apply the Pygmalion effect as a management strategy. If the empirical evidence primarily reflects the effects of differential expectations, it cannot be assumed that adopting high expectations across an entire workforce would produce comparable outcomes.

7. Future Research

These limitations highlight several important directions for future research. Greater attention should be devoted to investigating the intentional use of the Pygmalion effect within workplace settings to establish whether its psychological, organisational, and ethical consequences differ from those observed in educational contexts.

Future studies should also distinguish between the effects of high expectations and differential expectations in order to determine whether the outcomes associated with the Pygmalion effect are primarily driven by elevated expectations or by differential expectations.

Longitudinal research would further improve understanding of whether artificially induced expectations produce sustained performance improvements or contribute to burnout, anxiety, and declining psychological well-being over time. Addressing these gaps would contribute to a more comprehensive understanding of whether the intentional induction of the Pygmalion effect can be implemented as an effective and ethically responsible management strategy within 21st-century organisations.

8. Acknowledgements

I would like to express my sincere gratitude to my mentor, Sophie Dixon, for her guidance and patience throughout the making of this paper. I would also like to extend my heartfelt thanks to my parents, brother, and my loved ones for their moral support and encouragement. I am especially grateful to my close friends, whose constant encouragement helped me complete this paper.

Work Cited

Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. Nature Reviews Neuroscience, 10(6), 410-422. https://doi.org/10.1038/nrn2648

Bakker, A. B., & Demerouti, E. (2007). The job demands-resources model: State of the art. Journal of Managerial Psychology, 22(3), 309-328. https://doi.org/10.1108/02683940710733115

Baldissarri, C., Andrighetto, L., & Volpato, C. (2014). When work does not ennoble man: Psychological consequences of working objectification. TPM: Testing, Psychometrics, Methodology in Applied Psychology, 21(3), 327-339.

Brophy, J. E., & Good, T. L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. Journal of Educational Psychology, 61(5), 365-374. https://doi.org/10.1037/h0029908

Chen, J., Cooper-Thomas, H. D., & Cheung, G. (2023). Cue consistency matters: How and when newcomers respond to supervisor creativity expectations. The International Journal of Human Resource Management, 34(22), 4291-4312. https://doi.org/10.1080/09585192.2023.2189022

Cooper, H. M., & Hazelrigg, P. (1988). Personality moderators of interpersonal expectancy effects: An integrative research review. Journal of Personality and Social Psychology, 55(6), 937-949.
https://doi.org/10.1037/0022-3514.55.6.937

Crocker, J., & Wolfe, C. T. (2001). Contingencies of self-worth. Psychological Review, 108(3), 593-623. https://doi.org/10.1037/0033-295X.108.3.593

Eden, D. (1992). Leadership and expectations: Pygmalion effects and other self-fulfilling prophecies in organizations. The Leadership Quarterly, 3(4), 271–305. https://doi.org/10.1016/1048-9843(92)90018-B

Gruenfeld, D. H., Inesi, M. E., Magee, J. C., & Galinsky, A. D. (2008). Power and the objectification of social targets. Journal of Personality and Social Psychology, 95(1), 111-127. https://doi.org/10.1037/0022-3514.95.1.111

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. Proceedings of the National Academy of Sciences, 93(24), 13473-13480. https://doi.org/10.1073/pnas.93.24.13473

Good, T. L., Sterzinger, N., & Lavigne, A. (2018). Expectation effects: Pygmalion and the initial 20 years of research. Educational Research and Evaluation, 24(3–5), 99–123. https://doi.org/10.1080/13803611.2018.1548817

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. Journal of Social Issues, 57(4), 657-674. https://doi.org/10.1111/0022-4537.00234

Ichilov, A. (2025). The integration of the peace state and Pygmalion effect. Gaia, 1(1), 14-22.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. Personality and Social Psychology Review, 9(2), 131-155. https://doi.org/10.1207/s15327957pspr0902_3

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24, 167-202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mitchell, M. S., Greenbaum, R. L., Vogel, R. M., Mawritz, M. B., & Keating, D. J. (2019). Can you handle the pressure? The effect of performance pressure on stress appraisals, self-regulation, and behavior. Academy of Management Journal, 62(2), 531-552. https://doi.org/10.5465/amj.2016.0646

Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. Journal of Management, 46(1), 100-126. https://doi.org/10.1177/0149206319869435

Ryan, R. M., & Deci, E. L. (2017). Self-determination theory: Basic psychological needs in motivation, development, and wellness. Guilford Press.

Timmermans, A. C., & Rubie-Davies, C. M. (2018). Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? Relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. Educational Research and Evaluation, 24(3–5), 241–263. https://doi.org/10.1080/13803611.2018.1550837

Sutton, C. D., & Woodman, R. W. (1989). Pygmalion goes to work: The effects of supervisor expectations in a retail setting. Journal of Applied Psychology, 74(6), 943-950. https://doi.org/10.1037/0021-9010.74.6.943

Twenge, J. M. (2017). iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy, and completely unprepared for adulthood. Atria Books.

Weinstein, R. S. (2018). Pygmalion at 50: Harnessing its power and application in schooling. Educational Research and Evaluation, 24(3–5), 346–365. https://doi.org/10.1080/13803611.2018.1550842

Whiteley, P., Sy, T., & Johnson, S. K. (2012). Leaders' conceptions of followers: Implications for naturally occurring Pygmalion effects. The Leadership Quarterly, 23(5), 822-834. https://doi.org/10.1016/j.leaqua.2012.03.006

Xu, X., Wang, Y., Li, M., & Kwan, H. K. (2021). Paradoxical effects of performance pressure on employees' in-role behaviors: An approach/avoidance model. Frontiers in Psychology, 12, 744404. https://doi.org/10.3389/fpsyg.2021.744404

Unraveling the Problem: How Fabric Design Can Help End Microplastic Pollution
By Savera Dhaliwal

Abstract

Microplastic pollution from textiles is a major environmental issue, as fibers released during everyday activities like washing clothes contaminate water, ecosystems, and the food chain. Because clothing is used daily, the environmental impact is often overlooked, even though textile fibers are a significant source of microplastics. Understanding textile materials and techniques can help reduce microplastic release. This project aims to identify which cutting techniques cause microfiber release and which textiles, in general, release more microfibers. It has already been discovered that many microplastics and microfibers are released during laundry machine use and, due to this discovery, experiments were conducted using a washing machine. After washing the fabrics, six samples of each fabric were prepared: two using a fabric scissor, two using a rotary cutter, and two using a hot knife. After each fabric was cut using its respective cutting technique, the fabric was washed in the machines, and the water was drained, collected, and measured using a spectrometer. The natural fabric that released the most microfibers was 100% cotton, and the synthetic was recycled polyester. The rotary cutter produced the highest overall fiber released, while the hot knife produced the least. The findings of this project show that fabric release not only depends on the fabric but also the cutting method; when a hot knife is used across all materials, the number of microfibers decreases. Microplastics harm ecosystems and will continue to do so unless a solution is found that will either decrease the amount of microplastic released or completely eradicate the microplastics.

Introduction

Microplastics pollution is detrimental to the environment and has been contaminating ecosystems, food, and especially water. Fibers in textiles release these microplastics into water through systems such as washing machines. This topic is important because it is part of humans' everyday life, as clothing is worn everyday and its environmental impact is overlooked even though it affects the environment everyday. Researching fabrics can help determine what fabrics release the most microplastics as well as the treatments and methods of production and physical properties that make materials more prone to shedding which can make a solution that will reduce the number of microplastics released.

Researchers at environmental science, textile engineers, the National Institute of Health, and the EEA are all researching microplastics in textile currently. Microbiologists are also studying solutions to degrade the fibers and dyes using bacterias. The impact is global and affects ecosystems, the environment, humans, and animals. Communities that are located near textile production and have limited wastewater treatment are especially affected by this pollution. There are debates about which materials shed the most and specifically the debate about if recycled polyester shed more or virgin polyester is frequently mentioned. By finding out what techniques are used, treatments, and fabrics typically release the most microfibers, the effects of

textile pollution can be mitigated when textiles can be produced that will have the least amount of microplastics.

Countries with large textile industries and production as well as countries that have bad wastewater filtration systems and high population density are most affected by textile pollution. Textile manufacturing is extensive in South and East Asia, and therefore in those areas they are the most affected by microplastic pollution from textiles. This is an ongoing issue and has been more extensively studied over the last two decades. New discoveries are currently being made such as why some materials release more microfibers, what treatments and production factors cause more to be released, and some solution to degrade the chemicals, dyes, and fibers that are released. This topic connects with environmental science and textile engineering as it delves into what components of the textiles causes this shedding and the effect it has on the environment. All of these ideas combined show how important it is to study microplastic pollution from textiles further as it harms the environment and the health of humans and ecosystems extensively and commonly. A small change in the cutting techniques and fabric used may reduce the number of microfibers released into the environment and decrease the impact of pollution on the environment.

Microplastics end up in the environment through textiles as some textiles are made from plastic fibers. These fibers release into the environment through shedding, washing machines, and other factors such as friction. They then pollute water and can also enter the human system through ingestion and inhalation as the microplastics in fabrics shed and are then released into the environment. Additionally, fish can ingest these microplastics which then enter the human food chain. Research has noted that microplastics harm fish's gut, liver, kidneys, and endocrine receptors (Subaramaniyam, 2023). As fast fashion is still prevalent and global consumption rises, the cycle of purchasing clothes, washing them continuously, and then throwing them away causes microplastics to infect our environment and surroundings at a fast rate. This is a significant environmental concern that could be stopped by the textile industry if they implement better ways to make textiles that are eco-friendly. While environmental scientists focus on the contamination from the microplastics in the textiles, engineering adds a different layer to the research as it emphasizes the structural causes of the release. Many fabrics contain dyes and absorb chemicals which lead to textiles not only polluting the water but also poisoning the water and the environment with chemicals that are usually hard to degrade. Vujić et al. (2025) talks about the effect of microplastics on the environment as many of the textiles and their fibers absorb harmful pollutants like 4-MBC which is an endocrine disruptor and can contaminate water and ecosystems. Anwar et al. (2023) builds onto this as he talks about how dyes that are also absorbed by textiles contribute to harmful effects on the environment, such as stunting soil and plant growth. He experiments with bacteria to break down these dyes which shows that some dyes can be broken down by specific bacteria strains. This information by Anwar et al. (2023) can insure that dyes that can be broken down are used in these textiles to make it more eco-friendly and mitigate the effects of the pollution and the harmfulness of the textiles. Vujic adds more knowledge about what factors cause fabrics to be able to absorb these chemicals.

Using the information provided by Vujic et al. (2025), it is revealed that the amount of chemical that fabrics absorb can be reduced by looking at the polymer composition, the type of material, and fiber morphology, the physical structure of the fabric when trying to make a textile that is eco-friendly and won't absorb harmful chemicals. These two sources combined give lots of input about the effects that textiles have on the environment and the extent to the problems they are causing as they are containing their surroundings with the chemicals and dyes they absorb. Textiles are causing lots of problems in the world but this problem can be reduced by finding out what causes the release of microplastics and the science behind the structure and properties of textiles that make some more prone to shedding and more absorptive than others.

While these authors focus on the chemical impacts and variables that contaminate water, looking at what causes some fabrics to release microplastics, it is evident that it starts from the making of the textile and what techniques are used when making the material to affect the number of microplastics released. Textiles are mass produced with the industries focusing on efficiency and being able to produce textile at a fast rate rather than focusing on the how it will affect the environments which lead to these manufacturers to overlook the structural integrity of their products, leading to many fabrics being prone to shedding after minor abrasion and stress on the fabric. Therefore, as the textile industry plays an important role in the making of these textiles, they also play a significant role in the release of the microplastics as the structure of textiles are key to trying to reduce the release of these microfibers. It was found that fiber shedding was positively correlated with thickness of fabrics treatments which means that treatments and abrasion can cause the loosening of the structure of the textile and the release of microfibers (Vassilenko, 2021*).* Knowing this correlation, it is important to choose treatment that will not loosen the structure of the textile such as silicon softener and durable press treatments and instead use treatments like enzyme hydrolysis that will not loosen the fibers as much (Allen, 2024*).* The article by Allen et al. (2024) talks about cutting techniques also impacting the release of microfibers as scissor-cut edges shed more fibers than laser-cut or heat-sealed edges. It is hypothesized that this occurs because scissor-cut edges leave the fibers more exposed while allowing them to shed more while heat-sealing or laser-cutting the edges does not expose the fibers and will not expose them which means no exposed fibers shedding. Technique and treatments after the making of the fabrics and the choosing of the materials used can cause many of the microfibers to release as it loses the strings and fibers in the fabrics allowing for the release of the microfibers eventually as they are not as tight and secure. A gap in this study, however, is the amount of textiles that were tested and therefore additional research is needed to see if this conclusion applies to most textiles.

Moreover, Choi et al. (2022) also showed that the fabrics chosen are crucial in clothing making as typically acrylic and loosely woven fabrics shed the most and therefore tightly woven fabrics that shed less should be used to decrease the number of microfibers and microplastics released. Similarly Cui and Xu (2022) proved this idea as they show that satin fabrics release more fibers than plain fabric because satin fabrics have fewer interlacing points. This shows that satin is loosely woven in comparison to plain fabric because plain has more interlacing points

and therefore will release less microfibers because it is more tightly woven. The textile industry contributes to the microplastic problem as the treatments and machinery used on the fabrics and textiles can affect the number of microplastics that can be released. With the right fabrics and the right techniques the number of microfibers released can be reduced to help decrease the amount of microplastics in the environment. Additionally, Gliaudelytė et al. (2024) emphasizes the importance of fiber type, yarn structure, weave pattern, and mechanical or chemical finishes. The article talks about how the microplastics release is not solely based on the fiber type but also the composition, finishes, and structure of textile. If the composition, finishes, and structure of the textile is made so that the fibers will not be released as much and the weaves are tighter, the amount of microfibers and microplastics released will go down significantly. Ultimately, these microplastics end up in ecosystems and the human body and have harmful effects on them such as oxidative and DNA damage which is why an alternative textile and better techniques when making textiles should be used that will help reduce the microplastics pollution problem.

Research on microplastic pollution due to textiles reveals that structure of the textile, washing conditions, absorption of chemicals, and the treatments on the textiles all play a role in this pollution and the release of microplastics in the environment. By understanding why and how it plays a role, scientists can develop ways to reduce this pollution. *Unwrap*, for example, talks about eco-friendly material and textiles and what they are typically used for as well as what they feel and look like. Compared to these materials, they also bring up materials that are not eco friendly and what they feel and look like as well as their side effects and impacts on people and the environment. Using this information, alternatives to these harmful materials could be found that look and feel the same as the fabrics and materials that contain microplastics. While *Unwrap* talks about the materials and introduces alternatives and eco-friendly fabrics, they don't discuss what properties of the material makes it dangerous such as the structure and density. The article also categorizes the material into synthetic and natural fabrics and does not talk about the structural variables which make these fabrics good and bad (*Unwrap*). When looking at this information from an engineering standpoint, it is important to know why and what causes these materials to be dangerous. Additionally, it is crucial to look at what is dangerous but also why, which is what this website lacks. These gaps that the website fails to mention can be filled with the previous articles and research. Combined, enough information is given to understand what materials are eco-friendly or synthetic and also what makes these materials and fabrics shed more than others.

Using this information and research, it is shown that the amount of microplastics that are released from textiles significantly contribute to the microplastic problem, but it can be reduced through tighter weaves, more interlacing points, enzyme finishing, yarn twists, dyes that can be degraded, and making a material that is low-shedding, does not absorb chemicals, and environmental and cost friendly. After understanding the factors of what causes the release of microplastics, this research will test out different materials, fabrics, treatments, and undergo trials of abrasion and in washing conditions to make a textile that does not release microplastics. This research also strives to provide information of what causes microplastics to be released so

that textile companies can use this information to better their products and textile and use better machinery and techniques to ultimately decrease the number of microplastics and microfibers released.

Research Question: Which textile releases the least microplastics during washing, and what conditions and factors cause microplastics to be released in textiles?
Hypothesis: Synthetic fabrics will release more microfibers than natural fabrics, and the hot knife tool will have the least amount of microfibers released.

Methods/Materials

Research Type The project collected both quantitative and qualitative data using a spectrometer to measure how many microfibers were released from each fabric. Quantitative data was collected through absorbance & transmittance values, allowing for precise comparison of microfiber release across materials. Qualitative observations were also recorded by examining the physical condition of the fabrics before and after washing. This included noting visible changes such as edge fraying, fiber loosening, and overall structural integrity. These visual characteristics were then compared with the spectrometer data to identify patterns.

Sources: The articles about textile engineering and the environmental impact of microplastics and microfibers were used. These sources provide background information such as what engineering methods are said to cause more microfibers to release and techniques and treatments that should be experimented with when trying to design a textile. Info was gathered from websites like *Unwrap,* which provided a summary of fabrics, their uses, and if they are eco-friendly or not.

Methods: First, each of the textiles were measured to make sure that they are of comparable dimensions. Second, the fabrics were cut into two 20.32cm by 25.4cm samples using each of the cutting techniques creating six samples for each type of fabric. Fabrics were then observed, and quantitative data was written down. Records of how they felt and looked were taken whether they were coarse, smooth, thick, thin, shiny, or dull. For example, hemp felt coarse, and thick while satin polyester was smooth, and shiny. In comparison, minky fleece was soft, and fuzzy. These structural differences later helped interpret patterns in fiber release. The independent variable was the fabric and cutting technique used, and the dependent variable was the number of microfibers released. The control condition was the washing condition which was 500ml of water with one drop of detergent on normal for 4 minutes. Establishing a control condition ensured that washing conditions, which were extraneous variables, across all samples were the same to assure that microfiber release differences were due to fabric type and cutting technique alone. Each sample was then washed, and the samples of the water the fabric is washed in was collected. In addition, the water was strained under a mesh wire to collect the fibers. Half a cuvette was filled with each individual sample water, and then the fibers collected

from that sample were added and measured in a spectrometer. The spectrometer measured transmittance and absorbance levels to then determine which fabric released the most microfibers and what cutting technique is the most and least effective in the release of microfiber. The lower absorbance indicated a lower microfiber release. High transmittance levels additionally meant that there were lower amounts of microfibers in the water.

Practicalities: The main challenge is making sure that everything is cleaned thoroughly before and after each fabric goes in the wash.

Risk and Safety: There are not many physical risks except maybe being exposed to detergates, dyes, and microfibers To ensure that this research was conducted safely, gloves, goggles, and masks were worn. There were also sharp tools like needles, scissors and sewing machines, but they were handled with care.

Implications: This research will help add onto the information about textiles as well as helping others understand which engineering methods and textiles release less microfibers. Using this information, industries and manufacturers can make and choose textiles that release the least amount of microfibers. Most research focuses on the comparison of synthetic and natural textiles and the environmental effects, but this research combines both as it tests how engineering methods affect the microfiber release. By testing both synthetic and eco friendly fibers under controlled conditions, this study will provide more knowledge about how to design low-shedding fabrics and showing environmentally friendly fabrics as well as trying to reduce the number of shedding on fabrics that are most commonly used like nylon and polyester. Microplastic pollution happens everyday, especially through textiles as washing clothes and fast fashion cause the release of many microplastics into our environments. Therefore, identifying what weaves, blends and treatments that may be able to mitigate the effect of this pollution are crucial and is what this project is trying to find to create textiles that release less microfibers and microplastics.

Results

To understand how fabric type and cutting techniques affected microfiber release during washing, experiments were conducted and analyzed using a spectrometer. Once data was collected and levels were measured, comparisons were able to be made and clear patterns emerged. The samples cut by the rotary cutter had the highest amounts of microfiber release, the heat sealer released the least amount, while the scissor laid somewhat in the middle. The hot knife tool releasing the lowest amounts of microfibers confirmed the original hypothesis that the hot knife would be the most efficient. However, the second part of the hypothesis about the cotton releasing the least amount of microfibers was disproved. When cotton microfiber levels were measured, they had the highest average amounts of microfibers released. This challenged the original assumption that synthetic fabrics would release the most fibers. Furthermore, it was

found that out of the synthetic textiles, recycled polyester released the most amount of microfibers while satin polyester released the least amount. The number of microfibers released by cotton was higher than the amount released by recycled polyester.

Observations: Hemp feels coarse, not very soft, looks dull and dry, thick. 100% cotton is coarse and not soft as well. Hemp is a little more coarse and looks dry and dull. Stain polyester is smooth, shiny, thin while nylon spandex is smooth and not rough but thick. 100% polyester was very rough, coarse, and very thick. Minky fleece is fuzzy and soft compared to spandex jerseys which are thin and smooth. 100% rayon is thin and smooth but did not look very breathable as the material was very dense. Lastly, the poly cotton blend also does not look very breathable but is on the thinner side and pretty smooth.

Fabrics	Hot Knife	Rotary Cutter	Scissors
100% polyester	A:0.006 T:98	A:0.012 T:97	A:0.033 T:92.0
35% cotton 65% polyester	A:0.130 T: 74.1	A:0.024 T:94.5	A:0.017 T:96.7
100% Hemp Canvas	A:0.032 T:94.3	A;0.004 T:98.6	A:0.029 T:93.3
80% Nylon 20% Spandex	A:0.047 T:89.6	A:0.012 T:97.5	A:0.013 T:97.2
100% Rayon from Bamboo	A: 0.026 T:95.3	A:0.025 T:94.6	A:0.022 T:96.0
Jersey Polyester	A:0.071 T:85.1	A:0.013 T:97.0	A:0.016 T:96.2
Recycled Polyester	A:0.016 T:96.2	A:0.005 T:98.4	A:0.07 T:98.3
Minky Fleece	A:0.016 T:96.4	A:0.011 T:97.8	A:0.032 T:91.9
Satin Polyester	A:0.013 T:96.9	A:0.004 T:99.0	A; 0.013 T:97
100% cotton	A:0.071 T:94.6	A:0.22 T:95.5	A:0.070 T: 84.9

Table 1: Absorbance and transmittance values by fabric type and cutting techniques. Low absorbance indicates reduced microfiber release and low transmittance means that there were more microfibers in the water.

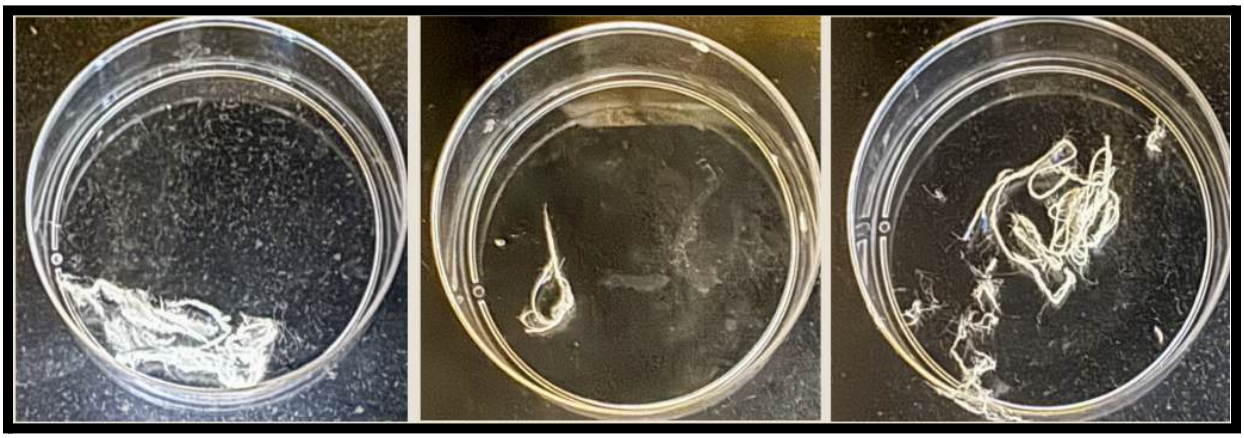


Figure 1: There are three petri dishes containing fibers that were collected from 3 different 100% cotton using the three different cutting techniques. On the left scissors were used, in the middle the heat sealer was used, and on the left the rotary cutter was used.

Discussion

A possible explanation to the cotton releasing the most amount of fibers is that cotton fibers, although natural, may be structurally weaker under agitation and more prone to mechanical breakage. In comparison, synthetic fibers, like polyester, are composed of long-continuous filaments that resist fragmentation. This aligns with previous research suggesting that fiber morphology and length influence microfiber shedding (Choi et al., 2022). While cotton released the most amount of microfibers on average, it is important to note that cotton is biodegradable. This means that the microfibers that it releases could absorb chemicals and harm, pollute, and affect the water and environment but would eventually degrade. In contrast, microfibers from synthetic textiles are microplastics that persist and can not decompose into the environment. This distinction is significant when considering environmental impact.

Testing the different cutting techniques shows that there are certain techniques that release more microfibers than others. Changing the type of cutting techniques, and using a hot knife, could reduce the amount of microfibers that are released and eventually mitigate the effects of microplastic pollution. A small, simple change in the textile industries has the potential to have large scale environmental benefits.

Through this research, it is proven that additionally the structure of the material and type of materials affect the amount of microfibers released as well. Material with observable looser weaves, such as cotton and polyester, released more microfibers than materials with tighter weaves, such as spandex, which was determined by connecting the qualitative data with the quantitative data.

Limitations The sample size could be expanded, and the spectrometer measures total suspended particles but does not differentiate between polymer types. The short washing duration may not fully represent household washing conditions, and a limited number of trials may reduce the statistical reliability of the results. Future research could incorporate microscopy or polymer-specific analysis techniques to identify the exact composition of released fibers. Further studies could also test full washing machine cycles, investigate enzyme treatments, or study weave density as a variable.

Conclusion

Instead of developing entirely new materials, it was investigated whether a small process change, something as simple as switching cutting techniques, could reduce environmental impact. This offers a practical and scalable perspective that could realistically be implemented in the textile industry. This research shows that heat sealing fabric edges reduces microfiber release, while rotary cutting increases shedding. Furthermore, fabric structure plays a significant role in fiber loss, with cotton releasing a surprisingly high quantity of fibers in this experiment. Microplastic pollution is a global issue, but solutions do not always need to be complex. These findings could be applied in textile manufacturing by encouraging the use of heat-sealing techniques during production. Additionally, consumers and companies could make informed

decisions about fabric selection to reduce environmental impact. If manufacturers adopted cutting methods that reduce fiber shedding before clothing even reaches consumers, the environmental impact could be reduced at the source. Small changes in production processes have the potential to create large-scale environmental benefits and help mitigate the effects of microplastic pollution. Given the scale of global textile production, even minor reductions in microfiber release per garment could collectively prevent billions of microplastic fibers from entering aquatic ecosystems each year. Microplastic pollution is known to affect aquatic animals by harming their endocrine receptors which hurts their reproduction system. If this pollution increases, it could affect humans similarly as well which is why solutions need to be found to decrease this pollution and prevent further harms to our ecosystems.

Works Cited

Unwrap. (n.d.). A plastic free guide to materials & fabrics. https://www.unwraplife.co/glossary?scrlybrkr=1e0cbc5c

Allen, E., Asuquo, E., Garforth, A., & Henninger, C. E. H. E. (n.d.). Microfiber pollution: A systematic literature review to overcome the complexities in knit design to create solutions for Knit Fabrics | Environmental Science & Technology. ACB Publications. https://pubs.acs.org/doi/10.1021/acs.est.3c05955

Anwar, M. O., Saleem, M., & Yahya, S. (2023). Synergistic Microbial Degradation of Microplastics and Toxic Dyes Showing Potential Reuse of the Degraded Dye Metabolites. Biology Bulletin, 50(4), 575–585. https://doi.org/10.1134/S1062359022601392

Choi, S., Kim, J., & Kwon, M. (2022). The Effect of the Physical and Chemical Properties of Synthetic Fabrics on the Release of Microplastics during Washing and Drying. Polymers (20734360), 14(16), 3384. https://doi.org/10.3390/polym14163384

Cui, H., & Xu, C. (2022). Study on the Relationship between Textile Microplastics Shedding and Fabric Structure. Polymers (20734360), 14(23), 5309. https://doi.org/10.3390/polym14235309

Gliaudelytė, U., Persson, M., & Daukantienė, V. (2024). Impact of textile composition, structure, and treatment on microplastic release during washing: A review. Environmental Science & Technology, 58(4), 1234–1245. https://journals.sagepub.com/doi/10.1177/00405175241260066

Subaramaniyam U, Allimuthu RS, Vappu S, Ramalingam D, Balan R, Paital B, Panda N, Rath PK, Ramalingam N and Sahoo DK (2023) Effects of microplastics, pesticides and nano-materials on fish health, oxidative stress and antioxidant defense mechanism. Front. Physiol. 14:1217666. doi: 10.3389/fphys.2023.1217666

Vassilenko, E., Watkins, M., Chastain, S., Mertens, J., Posacka, A. M., Patankar, S., & Ross, P. S. (2021). Domestic laundry and microfiber pollution: Exploring fiber shedding from consumer apparel textiles. PloS One, 16(7), e0250346. https://doi.org/10.1371/journal.pone.0250346

Vujić, M., Srebro, T. M., Vasiljević, S., Simetić, T., Molnar Jazić, J., Agbaba, J., & Tubić, A. (2025). Environmental Fate of 4-Methylbenzylidene Camphor: Adsorption Behavior on Textile-Derived Microplastic Fibers in Wastewater and Surface Water Systems. Materials (1996-1944), 18(16), 3799. https://doi.org/10.3390/ma181637

Acknowledgements

I would like to thank my sister, mom, and dad. My sister for introducing me to the science research class and helping me along the way with the obstacles I faced to propel me forward. My mom for taking countless hours listening to me spew on and on about my research and presenting my topic and poster to her. Lastly, my dad for giving me advice about research, medicine, and life in general.

Reassessing the Safe-Haven Premium: A Regression-Based Evaluation of U.S. Treasury Convenience Yield Underperformance Relative to Stock–Bond Correlation

By Taimur Lodhi

Abstract

This paper empirically investigates the erosion of the U.S. Treasury safe-haven status through the lens of the convenience yield and its historical relationship with stock-bond correlation. While prior literature has qualitatively assessed the perceived erosion of this safe-haven appeal amid fiscal concerns and the April 2025 tariff war shock, a quantitative measure of this erosion remains limited. Using daily U.S. equity and Treasury market data, this study constructs a PCA-derived convenience-yield proxy and regresses it on the rolling 60-day stock-bond correlation to establish a historical benchmark relationship. Regression residuals are then analyzed to identify statistically significant deviations from this benchmark. The results show that between May and September 2025, the convenience yield consistently underperformed its expected value by more than two regression standard errors, indicating significant underperformance relative to its historical benchmark. Even after controlling for inflation expectations using the 5y5y TIPS break-even rate, statistically significant negative residuals persist over a longer period. These findings provide evidence of a weakening in Treasuries' safe-haven status, suggesting that their historical hedging properties may have deteriorated during periods of macroeconomic stress. The results have important implications for portfolio risk management, asset allocation, and U.S. fiscal sustainability and investor confidence globally.

Keywords: U.S. Treasury, Convenience Yield, Stock-Bond Correlation, Underperformance, Safe-Haven Assets.

Introduction

A central part of the U.S. economy is the offering of its sovereign bonds: U.S. Treasuries. Treasuries are backed by the full faith and credit of the U.S. government (Simmons 325) and are therefore widely regarded as extremely safe investments.

One reason these assets are so attractive is their safe-haven status. This status refers to Treasuries' reputation as low-risk, highly liquid assets that investors flock to during periods of market uncertainty (Hager). However, in the past year, investors have noticed irregularities in U.S. Treasury yields. Specifically, the traditional negative stock–bond correlation has weakened, with Treasury and stock returns sometimes moving together instead of in opposite directions (Acharya and Laarits 1). While prior commentary has focused on qualitative indicators, this paper aims to empirically investigate this erosion of safe-haven status. The analysis is conducted through the lens of the convenience yield and its relationship with the stock–bond correlation. This is significant because U.S. Treasuries form the backbone of global portfolios and risk

Singapore American School, Singapore
Email: lodhi43185@sas.edu.sg
Corresponding author: Taimur Lodhi (sole author)

management. Any structural change in their safe-haven status, especially in their hedging value, has consequences that extend far beyond the U.S. Understanding whether this erosion is empirically measurable provides relevance for how trillions of dollars in global capital are allocated and protected during periods of market stress.

Historical Context

The tariff war shock that occurred in early April is considered one of the key drivers of the diminished safe-haven status of the U.S. Treasury, especially at the long end of the curve (Acharya and Laarits 1). On February 4th, 2025, the U.S. announced a 10% additional tariff on all Chinese imports. This caused a back-and-forth escalation, as China responded with retaliatory tariffs and the U.S. countered with further increases, resulting in the large-scale tariff war of early April 2025 that disrupted global financial markets (Feingold and Botwright). The result of tariffs is often stagflation, as it raises prices while disrupting growth (Acharya and Laarits 9), which can make Treasuries less effective hedges to other assets like stocks. A consequence of this is that the relative demand for U.S. Treasuries would then weaken as their non-monetary benefits, such as hedging value, deteriorated during the tariff war.

Literature Review

Diminished safe-haven status

U.S. Treasuries play a central role in global financial stability and government financing. If Treasuries no longer function reliably as a safe-haven, rising government debt may become more expensive to finance, with broader implications for fiscal sustainability and the stability of global capital markets. In this context, investor confidence in Treasuries could be undermined by the current climate of uncertainty, such as the tariff war in 2025 and rising fiscal deficit in the United States, which have increased inflation and partially elevated interest rates (Thiagarajan; Bensaidan). Conversely, Roula Khalaf suggests that Treasuries' safe-haven status has diminished over the years and cannot be solely attributed to the sudden disruption during 2025 (Khalaf). In summary, these three sources acknowledge this diminished safe-haven status of Treasuries, but they don’t agree on why, or over what time period, or even whether it has merely eroded or disappeared completely (Bensaidan; Khalaf; Thiagarajan). While these studies explore the potential drivers behind the diminished safe-haven status of Treasuries, it remains unclear how this erosion has affected their historical function as a hedge, particularly through the lens of the convenience yield.

Convenience yield

The convenience yield is the non-monetary benefit of holding a U.S. Treasury. This benefit arises due to Treasury securities’ safety and liquidity relative to other assets, often reflected in a lower yield on Treasuries than would otherwise be the case (Cieslak et al.). That is, because of this non-monetary benefit and safety aspect, investors are willing to accept a lower

interest rate on the Treasury itself. Additionally, considering that the convenience yield is the result of the safe-haven status, the rise in perceived U.S. risk and the fall in the convenience yield of long-term U.S. Treasuries are essentially two sides of the same story; investors see the U.S. market as riskier, so the "special safety benefit" of holding U.S. government debt increases (Corsetti et al.). Together, both sources illustrate that the convenience yield reflects this safety premium, or a non-monetary benefit of Treasuries, which is influenced by the volatility in the U.S. Treasury market (Cieslak et al.; Corsetti et al.). Yet while both sources establish the convenience yield as the measure through which safe-haven status is reflected in Treasury pricing, neither examines the convenience yield and its expected relationship with stock-bond correlation, a primary component of the convenience yield itself.

Convenience Yield and Stock-Bond Correlation

The relationship between convenience yield and stock–bond correlation is well established as an inverse one (Acharya and Laarits 4). When stock–bond covariance is low or negative, Treasuries serve as an effective hedge against stocks' risk as the two assets move in different directions, raising the convenience yield. Conversely, when stock–bond covariance is positive, Treasuries lose their hedging appeal because now Treasuries and stocks move in similar directions, leading to a decline in the convenience yield (Acharya and Laarits 1). Building on this, Gnewuch notes a similar trend in European sovereign bonds. During demand-side recessions, a negative stock-bond correlation boosts convenience yields as bonds act as a reliable hedge. However, in supply-side inflationary periods, this premium declines and the assets are stripped of their diversification value, because bonds and stocks fall in tandem. Both sources describe that the value of bonds as a hedge depends on their correlation with equities, and that hedging value directly impacts the convenience yield (which is reflective of the safe-haven appeal of Treasuries).

How the diminished safe-haven status has been reflected in the relationship between convenience yields and stock–bond correlation remains a scarcely researched topic. The only large-scale study that is similar to this topic was a paper published by Acharya and Laarits, two NYU Economics professors. The paper uses ETF data collected throughout the trading day to compute daily stock-bond covariance, then decomposes Treasury yields into risk-free rate, credit spread, and convenience yield components to analyze how the April 2025 tariff shock affected the safe-asset hedging properties of U.S. Treasuries across the yield curve. Their findings suggest that the convenience yield is closely tied to Treasuries' effectiveness as a hedge against equity risk, thereby underpinning a significant component of their safe-haven status (Acharya and Laarits 1).

Gap

While the diminished safe-haven status of Treasuries has been recognised in the last few months, along with the relationship between the convenience yield and stock-bond correlation, little has been said in regard to the relationship between the two. In particular, the literature is

characterized by an absence of commentary on the fair valuation of the convenience yield *vis-a-vis* the stock-bond correlation. Related work shows that increases in perceived U.S. risk compress the convenience yield, and that the long-maturity Treasury premium has trended lower in the post-crisis era (Corsetti et al.; Du et al.). But neither groups of research estimate a covariance-implied fair value for the convenience yield itself. That is, given a level of stock-bond correlation, what is the level of convenience yield suggested by the historical relationship of the two, and where is this level relative to prevailing estimates of the convenience yield? Finally, if the observed convenience yield is less than the model-implied fair value, this will constitute evidence that the safe-haven status of Treasuries has diminished. This investigation is an important step to quantitatively highlight the diminished safe-haven status through the means of analyzing the performance of the convenience yield with respect to the stock-bond correlation. Addressing this empirical gap helps clarify whether investors can still rely on Treasuries as a hedge against equity risk, which is extremely timely as markets navigate ongoing trade tensions and elevated stagnation risk (Gattiker et al.). In this environment, where traditional assumptions regarding portfolio construction are being tested, understanding whether the convenience yield has structurally deviated from its historical benchmark will quantitatively clarify and extend the existing qualitative research on whether Treasuries are still safe-haven assets to invest in during periods of market stress.

This paper will address this question: To what extent has the convenience yield of Treasuries underperformed its historical relationship with stock–bond correlation, as measured through a regression-based correlational analysis? The method the paper will use is quantitative regression-based correlational analysis, in which the convenience yield is regressed against the rolling 60-day stock-bond correlation to estimate a historical benchmark relationship, and deviations from that benchmark are then evaluated through regression residuals. It is hypothesized that this diminished safe-haven status of the Treasuries in 2025 can be reflected in the convenience yield underperforming its historical relationship with the stock-bond correlation.

Methodology

Method Rationale

Based on the current literature on this topic, if the convenience yield underperforms its historical relationship with stock–bond correlation, then this provides empirical evidence of a diminished safe-haven status of Treasuries.

Quantitative Regression-Based Correlational Analysis

With these considerations in mind, it is important to break down and define the different components of the method. Quantitative analysis relies on observations backed by empirical evidence, and a “correlational analysis” is a method that measures the strength and direction of association between two variables (Baraka; Janse et al.). In the context of this study's inquiry, this refers to determining the relationship between the convenience yield and stock-bond

correlation. The “regression” generates predicted values of the convenience yield for given levels of stock–bond correlation, using historical data to establish a benchmark relationship. This regression-based correlational approach is the most appropriate method for this study because it allows for the estimation of a historical benchmark relationship and the identification of systematic deviations from that benchmark, which descriptive statistics or qualitative approaches cannot rigorously measure. The method further builds on the structure of Acharya and Laarits (8), who decompose Treasury yields into their component parts, which include the convenience yield, and then correlate the change in convenience yield to stock-bond covariance to show that the erosion of the convenience yield relates to a loss of Treasuries' hedging property. This study extends beyond that by estimating a historical benchmark relationship and then analyzing deviations through regression residuals, which is an approach reaffirmed by Diercks and Asnani, who similarly treat regression residuals as economically meaningful diagnostics for Treasuries.

Pilot Study

A pilot study was conducted to determine the feasibility of the regression-based method before the full study. In this preliminary analysis, the 10-year term premium (ACMTP10) was used as the dependent variable instead of the convenience yield, as it is a widely used measure of the non-expected rate component of Treasury yields (Diercks and Asnani). Therefore, this makes it a more practical proxy for testing the capability of the method. The results showed that the term premium overperformed its expected value throughout 2025, confirming that the regression method can detect systematic deviations from historical norms (See Appendix A). The pilot study also revealed that results are sensitive to the rolling window length, particularly shorter windows like the 30-day, which captures short-term regime shifts but introduces noise, while the longer, 60-day window helped smooth variation, therefore giving more stable and reliable results than the shorter windows. The transition from the initial usage of a 30-day to a 60-day is a refinement carried forward in the main study, where the same structure is applied to the convenience yield.

Data and Sample Selection

The full study consists of daily observations of U.S. equity and Treasury market data. These data were selected to capture the relationship between equities and Treasuries, therefore reflecting the UST’s hedging role, which is central to theories of convenience yield (Acharya and Laarits 2). Daily observations are used to capture short-term fluctuations in market relationships that would be obscured in lower-frequency data. The same period from which data is collected spans from January of 2010 to March of 2026; this time frame was purposefully selected to allow the model to estimate a historical relationship across varied data that have experienced various macroeconomic conditions (Cenci 227; Mitton 529). This ensures the historical benchmark that is calculated will be representative of diverse conditions, such as the post-2008 financial crisis, the COVID-19 pandemic, and the tariff war in 2025.

Data Sources and Processing

Treasury yields are obtained from the Federal Reserve Economic Data (FRED). Yields on AAA-rated investment-grade corporate bonds come from FRED, and maturity-matched interest rate swap rates are drawn from Refinitiv Datastream and Bloomberg. To capture asset market movements, SPDR S&P 500 ETF (SPY) prices and iShares 20+ Year Treasury Bond ETF (TLT) prices are used. All datasets are aligned by date (See Appendix B).

Dependent Variable: Convenience Yield Construction

The dependent variable, the convenience yield, cannot be directly observed; it is estimated indirectly. Two spreads are first constructed as the difference between yields of matched maturities: (1) the yield spread between Treasuries and AAA-rated corporate bonds of matching maturity, and (2) the yield spread between Treasuries and maturity-matched swap rates (See Appendix C). These spreads isolate Treasury-specific safety and liquidity advantages. A Principal Component Analysis (PCA), a statistical technique that reduces multiple correlated variables into a smaller set of components that retain most of the information in the data (Groth et al. 542), is then performed on these two spread series. This approach is validated by Acharya and Laarits (6), who similarly infer the convenience yield through market-based spreads targeting Treasury-specific safety and liquidity benefits. Specifically, they construct a synthetic nominal yield that isolates the convenience yield as the difference between the observed yield and the model-implied rate. This supports the use of PCA and spread-based proxies as established methods for quantifying this unobservable variable. The first principal component, which captures their shared movement, is extracted and used as the proxy for the convenience yield (Variable 1). This PCA-derived convenience-yield series is also what provides the current, observed convenience yield for 2025. Because the same PCA factors are applied across the entire sample, the 2025 values of the series represent today's actual convenience yield, which can then be directly compared to the model-implied 'fair value' from the regression to determine whether Treasuries are underperforming.

PCA Justification and Validity

The use of the PCA is therefore justified because it isolates the common variation among multiple spread measures (Treasury corporate and Treasury swap spreads) that all reflect this same unobservable factor. By extracting the first principal component, the analysis is then able to summarize the shared movement driven by Treasury-specific liquidity and safety benefits, filtering out noise unique to any single spread. This approach provides a statistically replicable method for quantifying the convenience yield (Acharya and Laarits 14) and ensures that the variable used in the regression accurately represents the phenomenon being studied. In doing so, the instrument maintains both validity, because it captures the intended economic concept, and reliability, since PCA decomposition yields stable factor structures when applied to the same input data.

Independent Variable: Stock–Bond Correlation

The independent variable is the rolling 60-day stock-bond correlation (Variable 2), which measures how closely stock and bond returns move together. It is calculated as the Pearson correlation coefficient between daily percentage returns of SPY and TLT over a moving 60-day window (Molenaar et al. 17). A low or negative correlation indicates strong safe-haven behavior by Treasuries, while a rising correlation suggests a weakening of that status.

Control Variable: Inflation Expectations

Furthermore, an additional control variable is included to account for inflation expectations, measured using the 5y5y TIPS breakeven rate (See Appendix D). This variable was identified by Acharya and Laarits (18) as an important factor that influences the convenience yield and its relationship with stock–bond correlation, particularly through its effect on the hedging properties of Treasuries, and its inclusion in this study was further supported following feedback from Matthias Gnewuch, an Economist at the European Stability Mechanism (See Appendix E). In essence, this is important, as it accounts for inflation-related effects on bond yields, allowing the model to get a more accurate estimate of the convenience yield.

Regression Model Specification

To estimate this relationship, the following regression model is specified:

$$\textit{Convenience Yield}_t = \alpha + \beta 1 \text{ x } \textit{Stock-Bond Correlation} + \beta 2 \cdot \pi t \quad + €t$$

Here, πt represents the 5-year/5-year forward inflation expectation rate, included as a control variable to account for inflation-driven variation in the convenience yield. εt represents the residuals, or in other words, the differences between the observed and predicted convenience yield given the stock-bond correlation. To determine whether Treasuries have recently underperformed their historical pattern, the rolling 520-day proportion of residuals exceeding one standard deviation is calculated. The 520-day window (approximately two trading years) was selected to capture a sufficiently long rolling horizon to differentiate persistent structural deviations from short-term noise, while also remaining sensitive enough to detect regime shifts within the 2025 out-of-sample period. This proportion measures how frequently the convenience yield significantly deviates from expectations. Negative residuals imply periods of underperformance, while positive residuals imply outperformance. While the model tracks deviations in both directions, this study's hypothesis specifically concerns negative residuals, as the claim is that safe-haven status has eroded, not strengthened.

Replicability and Ethics

All computations can be replicated using statistical software such as R, Python, or Excel, which support PCA, correlation, and regression functions (See Appendix F). The method directly answers the research question by identifying when and to what extent the convenience yield diverges from its expected relationship with the stock-bond correlation, thereby evaluating the

strength of Treasuries' safe-haven status. Since this study relies on financial market data obtained from established sources such as FRED, Bloomberg, and Refinitiv Datastream, no human subjects are involved, no private personal data are collected, and the research therefore does not require IRB review.

Seminal Source Connection

In *When Do Treasuries Earn the Convenience Yield? — A Hedging Perspective,* Acharya and Laarits examine how the convenience yield of Treasuries evolves in response to market shocks. While my inquiry is still based on the same theoretical structure, it deviates from Acharya and Laarits's paper in the sense that the measurement and empirical design are slightly tweaked, and this is because the research goals differ. Their paper is an event-driven analysis, focusing on intraday responses to a specific market shock, and they decompose yields to explain simultaneous changes in the hedging behaviour of Treasuries. On the other hand, this study is predominantly historical and comparative. While Acharya and Laarits (10) employ multiple market-based proxies to estimate the convenience yield, this study builds a PCA-based convenience yield measure using Treasury-corporate and Treasury-swaps spreads, which detects variations in safety and liquidity premia across a few near risk-free benchmarks. The choice of doing this is to focus on comparability over time and the identification of historical patterns and deviations, rather than short-term intraday dynamics. Methodologically, this study further differs from Acharya and Laarits's study as it estimates a regression-based historical benchmark between convenience yield and stock-bond correlation and evaluates the deviations through the form of analyzing regression residuals. These changes are justified because my study does not look to explain why a shock occurred or why a shift occurred, but instead to assess whether current convenience yield levels are unusually low given the historical market benchmark.

Limitations

A main limitation of this study is the indirect measurement of the convenience yield, which must be estimated with the use of a PCA-based proxy and may capture noise unrelated to true shifts in investor demand for safety or liquidity. Similarly, the rolling 60-day stock-bond correlation may smooth over short-term regime shifts, therefore missing some shorter periods of possible data changes. Another limitation is the assumption of a stable, linear relationship between the two variables, which may overlook nonlinear behavior, especially during crises when safe-haven dynamics behave differently. Finally, since the study is correlational, it cannot establish causation or isolate the effects of specific policy or macroeconomic drivers.

Results and Discussion

The hypothesis that the convenience yield would underperform its historical relationship with stock-bond correlation in late 2025 is supported by the following findings. More specifically, these findings provide direct empirical evidence for the hypothesis that the

diminished safe-haven status of U.S. Treasuries in 2025 would be reflected in a statistically significant underperformance of the convenience yield relative to its historical benchmark.

Baseline Regression Results

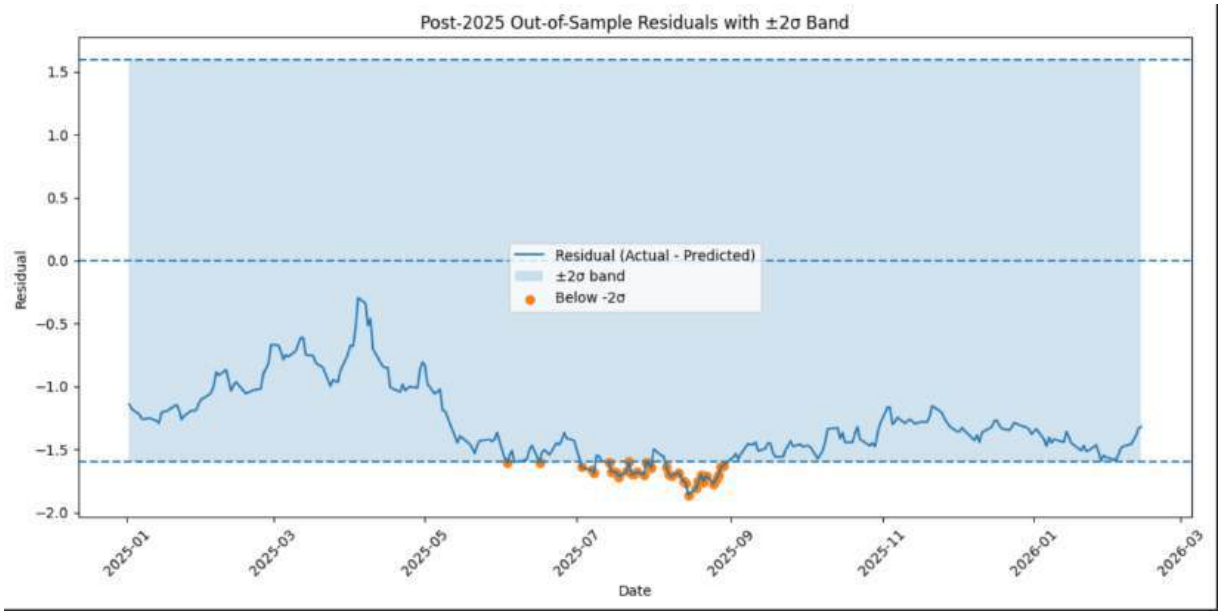


Fig 1: Stock-Bond Correlation and Convenience Yields — 2025 Out-of-Sample Residuals

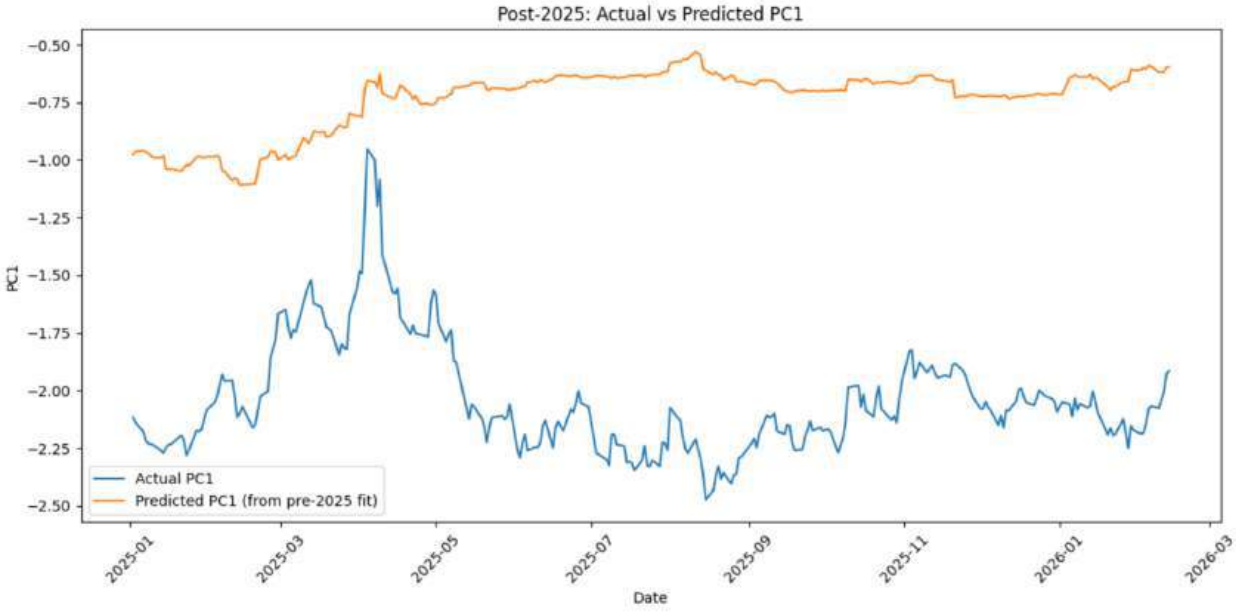


Fig 2: Out-of-Sample: Actual vs Predicted PC1

The regression analysis reveals a clear deviation in the relationship between the convenience yield and its historical determinants (Figure 2). While prior literature explains both the behaviour of the convenience yield through hedging demand and inflation expectations and

the appearance of a diminished safe-haven status, it doesn't account for potential breakdowns of these relationships over time, thereby lacking a true quantitative measurement proving this diminished safe-haven status of the U.S. Treasuries. Using a baseline model with the 60-day rolling stock-bond correlation as the sole independent variable, the model explains approximately 50 percent of the variation in the convenience yield ($R^2 = 0.499$; see Appendix G). This moderately strong relationship between Treasury hedging value and the convenience yield movements corroborates the relationship identified by Acharya and Laarits (4), who identified stock-bond covariance as a driver of the convenience yield; therefore, under normal conditions, Treasury hedging value reliably predicts the convenience yield (Acharya and Laarits 16). These results extend the literature by illustrating that the relationship between the convenience yield and stock-bond correlation can break down under periods of elevated macroeconomic stress.

Out-of-Sample Underperformance (2025)

Furthermore, when examining the out-of-sample period between May 2025 and September 2025, the actual convenience yield falls below predicted values by more than two standard errors of the regression, representing a statistically significant negative deviation (See Appendix H). This finding directly answers the research question. The convenience yield systematically underperformed its model-implied fair value during this period, constituting empirical evidence of a structural break in Treasuries' safe-haven status and not just a temporary fluctuation. Additionally, the orange dots in Figure 1 represent statistically significant residuals outside the 2 standard deviation range during this period. The negative residuals between this specific period likely coincide with the lingering effects of the April 2025 tariff war escalation and a period of intensified investor concern about the U.S. fiscal outlook (Gattiker et al.). Specifically, the tariff shocks introduced stagflation risk by raising costs while slowing economic growth, which weakened the negative stock–bond correlation and reduced Treasuries' effectiveness as a hedge, thereby lowering the convenience yield (Acharya and Laarits 5). Simultaneously, rising fiscal concerns increase Treasury supply and perceived risk, reducing their scarcity and safe-haven premium, which further compresses the convenience yield (Bensaidan; Thiagarajan). Therefore, this indicates that the periods of significant deviation seen in the results are in alignment with the current literature's qualitative assessments of when geopolitical tensions and macroeconomic events influenced the safe-haven status of the Treasury.

Extended Model with Inflation Expectations

To further test this convenience yield breakdown, the model is extended to include the 5y5y TIPS breakeven inflation rate as an additional explanatory variable, which increases explanatory power to approximately 65% ($R^2 = 0.648$; see Appendix I). This suggests that inflation expectations, as well as hedging demand, play a strong role in determining the convenience yield. Essentially, by incorporating inflation expectations alongside the hedging dynamics, the model more fully captures macroeconomic conditions influencing the convenience yield, serving as a better predictor of the convenience yield. This inclusion is further supported by Arcidiacono et al., who identify inflation expectations as a key determinant of bond yields

and use them as a control variable to confirm that observed effects are not driven by broader economic conditions. In essence, the inclusion of inflation expectations represents a step toward a more complete coverage of the factors that explain the convenience yield.

Persistence and Magnitude of Underperformance

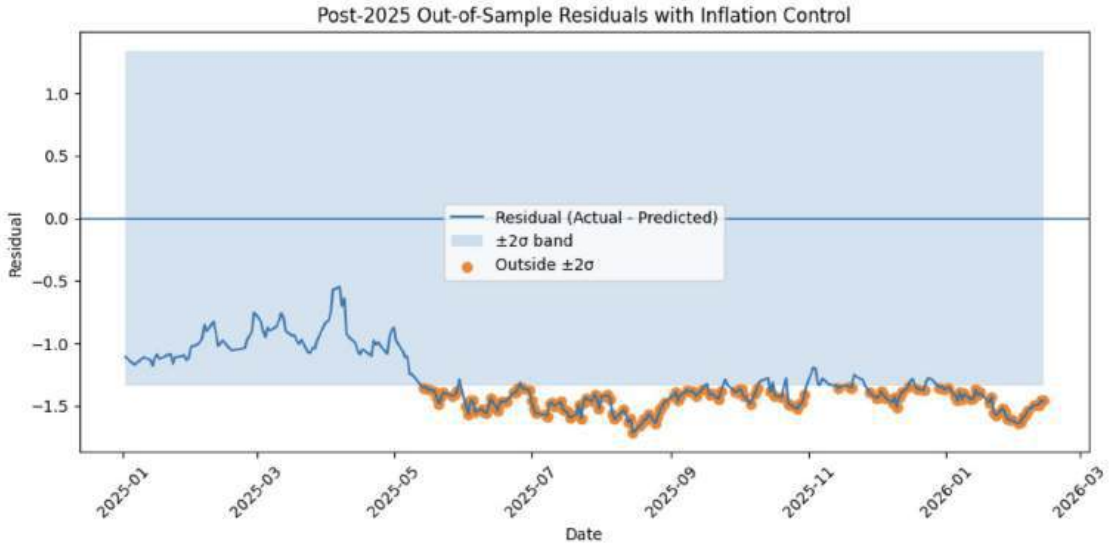


Fig 3: Stock-Bond Correlation, Inflation Expectation, and Convenience Yield — 2025 Out-of-Sample Residuals

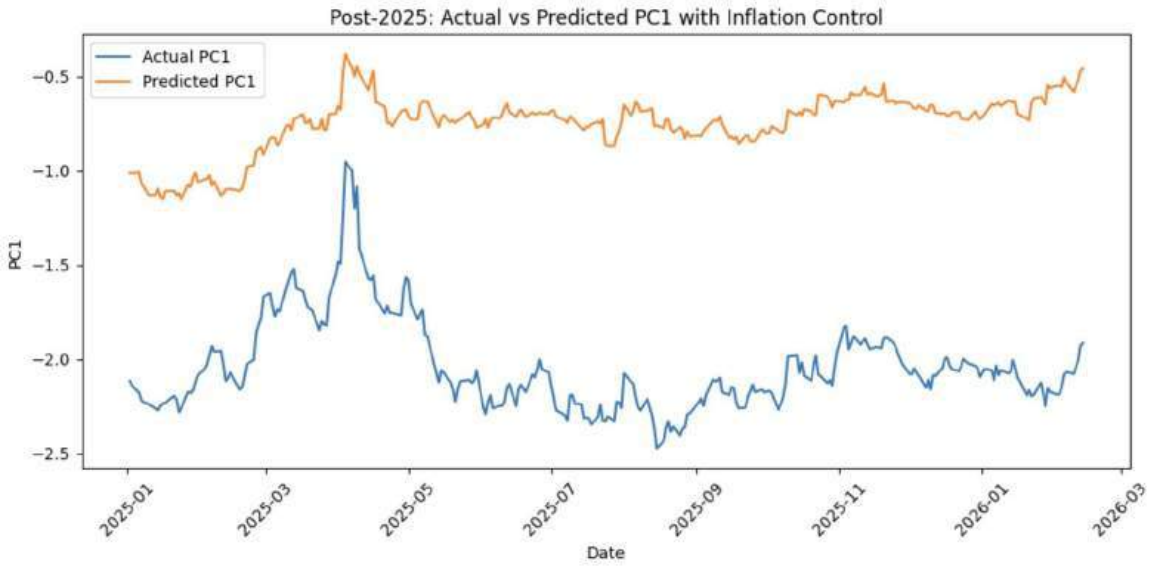


Fig 4: 2025 Out-of-Sample: Actual vs Predicted PC1

Moreover, the inclusion of the inflation expectations amplifies the observed deviation in the 2025 out-of-sample period. In Figure 4, we see a similar underperformance pattern as in Figure 2 from the first test. However, the duration of statistically significant underperformance as

seen in Figure 3 becomes longer compared to what's seen in Figure 1, and the magnitude of the residuals increases. This is critical because it signifies that even after accounting for both hedging demand and inflation expectations, the observed convenience yield remains persistently lower than predicted values, thereby suggesting that additional factors may be influencing the Treasury pricing beyond the scope of the model. This has important implications, as it suggests that Treasuries may no longer provide the same level of portfolio protection during periods of market stress, fundamentally changing their role in risk management and asset allocation strategies.

Interpretation of Findings

Overall, the results provide a quantitative answer to the research question: the convenience yield exhibited persistent, statistically significant underperformance relative to its model-implied fair value, even after controlling for both hedging value and inflation expectations. This extends the current literature by providing statistical evidence that the convenience yield of the U.S. Treasuries deviated meaningfully from its historical determinants in the 2025 out-of-sample period, suggesting a more structural shift rather than a temporary fluctuation.

Role of Unobserved Risk Factors

Even after controlling for both stock-bond correlation (hedging value) and inflation expectations, the significant negative residuals indicate that investors are requiring a higher-than-expected yield to hold Treasuries, since their safe-haven elements have weakened. However, it should be noted that since the convenience yield is not a specific numeric value, it must be measured indirectly. In the context of this study, it was measured through a PCA-derived proxy. The residuals capture the net effect of all unmodelled forces rather than that of one driver, meaning that the analysis identifies the magnitude of the deviation but doesn't pinpoint the precise cause.

Structural Shift in Treasury Pricing

The persistence and statistical significance of the negative residuals here would indicate the presence of a structural risk premium that extends beyond what inflation expectations and hedging dynamics can explain on their own. With this, Treasuries are no longer fully valued for their safety characteristics, because they are increasingly assessed through a lens of risk, and so concerns about debt and politics reduce their attractiveness as a risk-free asset. This aligns with the findings of Acharya and Laarits (9), who found that the April 2025 tariff war created stagflation risk that weakened the hedging value of Treasuries, which triggered investors to rotate toward other safe-haven assets like gold. This is a precise macroeconomic situation in which a structural risk premium would emerge, one where Treasuries lose their insurance-like properties across both inflation and growth shocks. This then leaves investors without the negative stock-bond correlation they have historically relied on; they demand this premium to compensate for this loss of a hedge. Unlike inflation expectations or hedging demand, these

broader macroeconomic and political risks are more difficult to quantify, but their effects are clearly illuminated in the persistent negative residuals.

Conclusion

Based on the findings of this paper, which documents statistically significant underperformance of the convenience yield relative to its historical relationship with stock-bond correlation, it is evident that the U.S. Treasury's safe-haven role has eroded in 2025. The convenience yield, as it stands, represents the quantitative measure for the safe-haven status, and an identifying factor of this convenience yield is the ability for Treasuries to provide a strong hedge (negative stock-bond correlation) towards other assets like equities. Therefore, by quantitatively measuring how the convenience yield has deviated from its historical benchmark with said hedging value, this study provides empirical evidence of a structural weakening of the safe-haven status of U.S. Treasuries. This study adds to the literature by providing the quantitative analysis that prior qualitative assessments of Treasury safe-haven erosion were missing, and confirms that the convenience yield's relationship with stock-bond correlation is not stable under macroeconomic stress.

Implications for Investors and Portfolio Theory

For institutional investors and risk managers, the most consequential implication concerns the validity of standard portfolio risk models. Value at risk models and mean-variance optimization layouts typically assume a negative stock and bond correlation. This is an assumption that has historically justified significant allocations to long Treasury bonds as a hedge against equity dips. However, as the results show a decline in the convenience yield, this assumption no longer holds. Additionally, considering the current political and economic state as of April 2026, this breakdown is occurring precisely during a period of market stress, when hedging is most needed. This means that institutional risk models built on historical correlation data are underestimating portfolio drawdown risk in these stress scenarios, which is a failure with direct consequences for capital adequacy and investor protection. Portfolio managers would therefore have to re-examine whether their hedging strategies, including the conventional 60/40 equities-to-fixed-income distribution, remain suitable and appropriate in a time where Treasuries’ safe-haven premium has permanently weakened.

Implications for Policymakers

For policymakers, the most significant implication of these findings relates to fiscal sustainability. If the structural risk premium identified in this study continues to be demanded, the U.S. government faces a situation where borrowing costs remain elevated even as inflation stabilizes, which is a form of fiscal dominance in which market confidence, instead of monetary conditions, becomes the binding constraint on debt sustainability. This suggests that restoring the safe-haven status may require a credible commitment to fiscal consolidation and not just monetary policy that rebuilds investor confidence in the long-run reliability of the Treasury.

Market Implications and Mean Reversion

From a market opportunity perspective, the increased risk premium also creates a potential mean reversion opportunity. A mean-reversion opportunity occurs when an asset is temporarily overvalued or undervalued relative to its typical level, and investors expect it to come back to its mean, which allows them to profit from that adjustment (Chen). If the deviation reflects temporary overpricing of risk rather than an actual permanent shift, compression of this premium would lead to declining yields and rising prices of USTs along the duration curve. However, the evidence from this study, most particularly the fact that the deviation persisted and deepened even after controlling for inflation expectations, is more consistent with an actual structural shift rather than one that is momentary, and investors therefore would weigh that interpretation in their decision with this strategy.

Scope and Limitations

The R^2 of 0.648 in the refined model is also worth pausing on here, because while it does account for a significant portion of variability in the convenience yield, the remaining 0.352 reflects a lack of certain explanatory variables. The unexplained variation is not just noise, as it likely reflects the very external macroeconomic influences the paper identifies but can't fully model, like fiscal policy expectations and trade tensions. Therefore, this limits the precision of the convenience yield estimate.

These limitations, however, do not invalidate the findings but instead define what this study can and cannot claim. For instance, the study can identify the magnitude and persistence of the convenience yield's deviation from its historical benchmark, but cannot establish causation or isolate a single macroeconomic driver that influences the convenience yield. Altogether, the study quantitatively demonstrates that the convenience yield has underperformed its historical relationship with stock-bond correlation in 2025, thereby serving as statistical evidence of the eroded safe-haven status of the U.S. Treasury bonds.

Future Research

The study's limitations beget new areas of research. Specifically, by addressing the proportion of variation in the convenience yield that is not explained, extended research incorporating variables that account for this would strengthen the causal interpretation of these findings. It is likely that the 35% of unaccounted variation comprises considerable macroeconomic drivers, which may influence the value of the convenience yield that is estimated and therefore the degree to which this quantitative assessment of a weakened safe-haven role can be generalised.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Disclosure statement

No potential conflict of interest was reported by the author.

Author contributions

The author is solely responsible for the conception, design, data collection, analysis, interpretation of results, and writing of this manuscript.

Data availability statement

The data that support the findings of this study are available from the respective data providers. Publicly available data were obtained from the Federal Reserve Economic Data (FRED) database, while Bloomberg and Refinitiv Datastream were used as data sources for financial market data. Derived data supporting the findings of this study are available from the corresponding author upon reasonable request.

Works Cited

Acharya, Viral V., and Toomas Laarits. "Tariff War Shock and the Convenience Yield of U.S. Treasuries – A Hedging Perspective." New York University Stern School of Business, 23 Apr. 2025, pages.stern.nyu.edu/~sternfin/vacharya/public_html/pdfs/Acharya_Laarits_tariff_202504.pdf.

---. "When Do Treasuries Earn the Convenience Yield? — A Hedging Perspective." New York University Stern School of Business, 16 Jan. 2025, w4.stern.nyu.edu/sternfin/vacharya/public_html/pdfs/working-papers/AcharyaLaarits_202501_v3.pdf.

Arcidiacono, C., et al. "Dangerous Liaisons? Debt Supply and Convenience Yield Spillovers in the Euro Area." CEPS, 15 Nov. 2024, www.ceps.eu/ceps-publications/dangerous-liaisons-debt-supply-and-convenience-yield-spillovers-in-the-euro-area/.

Baraka, S. "What Is Quantitative Analysis? Definition and Methods." Indeed, 16 Dec. 2025, www.indeed.com/career-advice/career-development/what-is-quantitative-analysis.

Bensaidan, A. "The U.S. Treasuries Market: An Idol with Feet of Clay. A Safe Haven Put to the Test." BNP Paribas Economic Research, 9 Dec. 2025, economic-research.bnpparibas.com/html/en-US/US-Treasuries-market-idol-with-feet-clay-safe-haven-test-9/12/2025,51826.

Cenci, S. "A Causal Approach to Test Empirical Capital Structure Regularities." The Journal of Finance and Data Science, vol. 8, 2022, pp. 214-28, doi.org/10.1016/j.jfds.2022.09.002.

Chen, J. "What Is Mean Reversion, and How Do Investors Use It?" Investopedia, 11 Dec. 2025, www.investopedia.com/terms/m/meanreversion.asp.

Cieslak, Anna, et al. "Inflation and Treasury Convenience." Becker Friedman Institute, 3 Sept. 2024, bfi.uchicago.edu/working-papers/inflation-and-treasury-convenience/?occurrence_id=0.

Corsetti, G., et al. "U.S. Risk and Treasury Convenience." Jane Way Institute, Cambridge, 16 Sept. 2025, www.econ.cam.ac.uk/sites/default/files/publication-cwpe-pdfs/cwpe2570.pdf.

Diercks, A. M., and D. Asnani. "The Treasury Tantrum of 2023." Board of Governors of the Federal Reserve System, 3 Sept. 2024, www.federalreserve.gov/econres/notes/feds-notes/the-treasury-tantrum-of-2023-20240903.html.

Du, W., et al. "The U.S. Treasury Premium." Journal of International Economics, vol. 112, 2018, pp. 167-81, doi.org/10.1016/j.jinteco.2018.01.001.

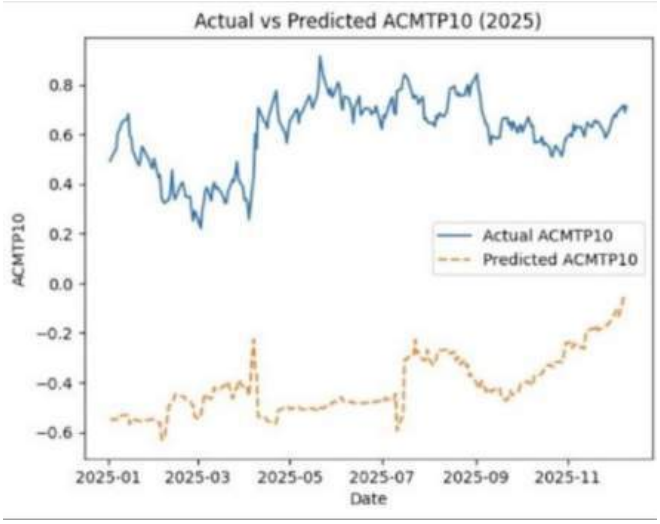
Feingold, Sarah, and Kate Botwright. "Tracking Tariffs: Key Moments in the US-China Trade Dispute." World Economic Forum, 25 June 2025, www.weforum.org/stories/2025/06/trumps-us-china-trade-tariffs-timeline/.

Gattiker, C., et al. "US-China Trade War Escalation: What Investors Need to Know." Julius Bär, 16 Apr. 2025,

www.juliusbaer.com/en/insights/market-insights/market-outlook/us-china-trade-war-escalation-what-investors-need-to-know/.
Gnewuch, M. "Sovereign Bonds, Convenience Yields, and the Resurgence of Supply Shocks." CEPR, 6 June 2025, cepr.org/voxeu/columns/sovereign-bonds-convenience-yields-and-resurgence-supply-shocks.
Groth, D., et al. "Principal Components Analysis." Methods in Molecular Biology (Clifton, N.J.), vol. 930, 2013, pp. 527-47, doi.org/10.1007/978-1-62703-059-5_22.
Hager, S. "A Global Bond: Explaining the Safe-Haven Status of U.S. Treasury Securities." EconStor Leibniz Information Centre for Economics (ZBW), 13 July 2016, www.econstor.eu/bitstream/10419/157799/1/bna-483_2016_hager_a_global_bond_preprint.pdf.
Janse, R. J., et al. "Conducting Correlation Analysis: Important Limitations and Pitfalls." Clinical Kidney Journal, vol. 14, no. 11, 2021, pp. 2332-37, doi.org/10.1093/ckj/sfab085.
Khalaf, Roula. "RIP the Convenience Yield of U.S. Treasuries? An Inconvenient Truth." Financial Times, 1 June 2025, www.ft.com/content/c2180890-f2c0-4240-9575-0348b9e3aa18.
Mitton, T. "Methodological Variation in Empirical Corporate Finance." The Review of Financial Studies, vol. 35, no. 2, 2021, pp. 527-75, doi.org/10.1093/rfs/hhab030.
Molenaar, R., et al. "Empirical Evidence on the Stock–Bond Correlation." Financial Analysts Journal, vol. 80, no. 3, 2024, pp. 17-36, doi.org/10.1080/0015198X.2024.2317333.
Simmons, E. C. "Treasury Deposits and Excess Reserves." Journal of Political Economy, vol. 48, no. 3, 1940, pp. 325-43. JSTOR, www.jstor.org/stable/1825591.
Thiagarajan, R. "The Great Repricing: Are U.S. Treasuries Still a Safe Haven?" State Street, July 2025, www.statestreet.com/hk/en/insights/the-great-repricing-us-treasuries.

APPENDICES

Appendix A: Pilot study results, which show the overperformance of the ACMTP10 compared to its predicted value.



Appendix B: Alignment of data based on dates. Spreadsheets consisting of the United States 10-Year Bond Yield Historical and SPY ETF Stock Price History.

United States 10-Year Bond Yield Historical

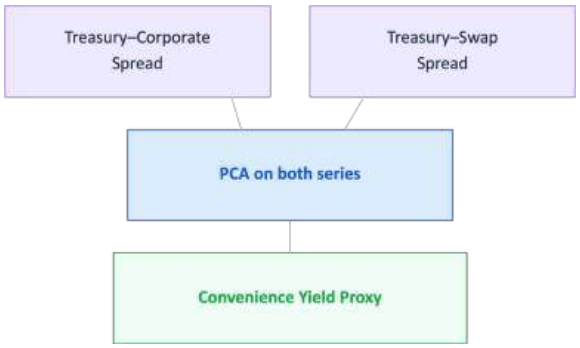
Date	Price	Open	High	Low	Change %
01/04/2010	3.823	3.823	3.823	3.823	-0.36%
01/05/2010	3.763	3.763	3.763	3.763	-1.57%
01/06/2010	3.829	3.829	3.829	3.829	1.75%
01/07/2010	3.827	3.827	3.827	3.827	-0.05%
01/08/2010	3.836	3.836	3.836	3.836	0.24%
01/11/2010	3.824	3.824	3.824	3.824	-0.31%
01/12/2010	3.717	3.717	3.717	3.717	-2.80%
01/13/2010	3.793	3.793	3.793	3.793	2.04%
01/14/2010	3.742	3.742	3.742	3.742	-1.34%
01/15/2010	3.676	3.676	3.676	3.676	-1.76%
01/19/2010	3.698	3.698	3.698	3.698	0.60%
01/20/2010	3.653	3.653	3.653	3.653	-1.23%

SPY ETF Stock Price History

Date	Price	Open	High	Low	Vol.	SPY_return
01/04/2010	113.33	112.37	113.39	111.51	118.94M	1.70%
01/05/2010	113.63	113.26	113.68	112.85	111.58M	0.26%
01/06/2010	113.71	113.52	113.99	113.43	116.07M	0.07%
01/07/2010	114.19	113.50	114.33	113.18	131.09M	0.42%
01/08/2010	114.57	113.89	114.62	113.66	126.40M	0.33%
01/11/2010	114.73	115.08	115.13	114.23	106.38M	0.14%
01/12/2010	113.66	113.97	114.21	113.22	163.33M	-0.93%
01/13/2010	114.62	113.95	114.94	113.37	161.82M	0.84%
01/14/2010	114.93	114.49	115.14	114.42	115.78M	0.27%
01/15/2010	113.64	114.73	114.84	113.20	212.28M	-1.12%
01/19/2010	115.06	113.62	115.13	113.59	139.17M	1.25%

[Screenshot: January 2010 - 2026 observations] *[Screenshot: January 2010 - 2026 observations]*

Appendix C: Simple visual mapping of the PCA formation for the convenience yield.



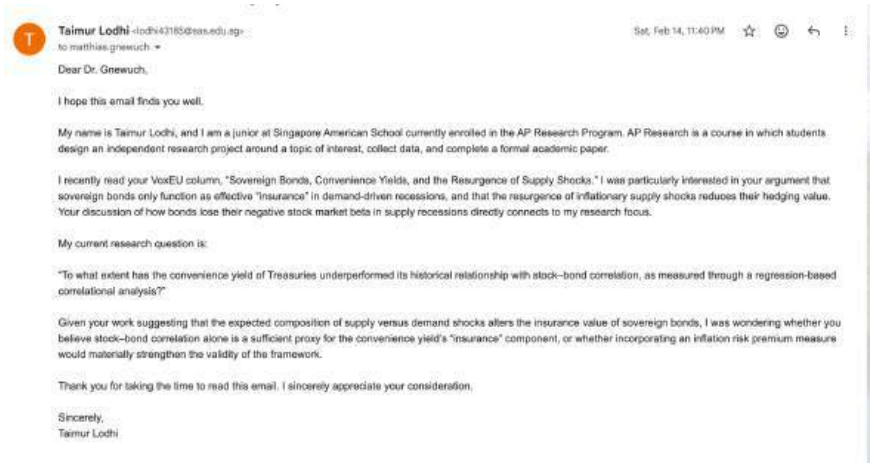
Appendix D: Inclusion of 5y5y TIPS breakeven rate from 2010 - 2026.

T5YIFR

Date	Inflation
2010-01-05	2.56
2010-01-06	2.58
2010-01-07	2.63
2010-01-08	2.66
2010-01-11	2.64
2010-01-12	2.63
2010-01-13	2.61
2010-01-14	2.60
2010-01-15	2.62

[Screenshot: January 2010 - 2026 observations]

Appendix E: Incorporation of Inflation Expectations Based on Feedback from Matthias Gnewuch



Taimur Lodhi <lodhi43185@sas.edu.sg> Sat, Feb 14, 11:40 PM

to matthias.gnewuch

Dear Dr. Gnewuch,

I hope this email finds you well.

My name is Taimur Lodhi, and I am a junior at Singapore American School currently enrolled in the AP Research Program. AP Research is a course in which students design an independent research project around a topic of interest, collect data, and complete a formal academic paper.

I recently read your VoxEU column, "Sovereign Bonds, Convenience Yields, and the Resurgence of Supply Shocks." I was particularly interested in your argument that sovereign bonds only function as effective "insurance" in demand-driven recessions, and that the resurgence of inflationary supply shocks reduces their hedging value. Your discussion of how bonds lose their negative stock market beta in supply recessions directly connects to my research focus.

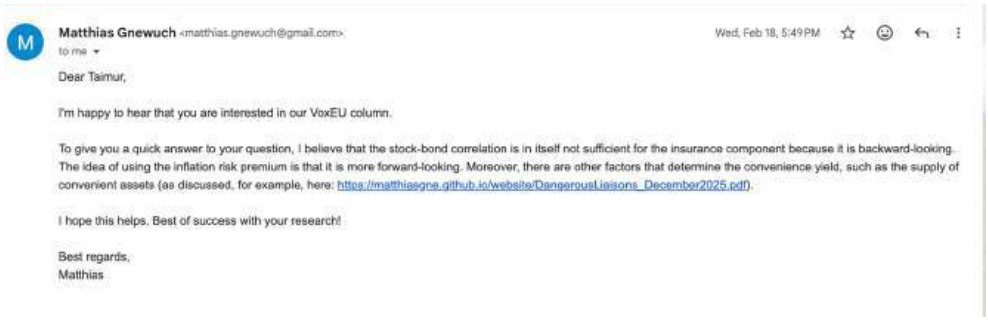
My current research question is:

"To what extent has the convenience yield of Treasuries underperformed its historical relationship with stock–bond correlation, as measured through a regression-based correlational analysis?"

Given your work suggesting that the expected composition of supply versus demand shocks alters the insurance value of sovereign bonds, I was wondering whether you believe stock–bond correlation alone is a sufficient proxy for the convenience yield's "insurance" component, or whether incorporating an inflation risk premium measure would materially strengthen the validity of the framework.

Thank you for taking the time to read this email. I sincerely appreciate your consideration.

Sincerely,
Taimur Lodhi



Matthias Gnewuch <matthias.gnewuch@gmail.com> Wed, Feb 18, 5:49 PM

to me

Dear Taimur,

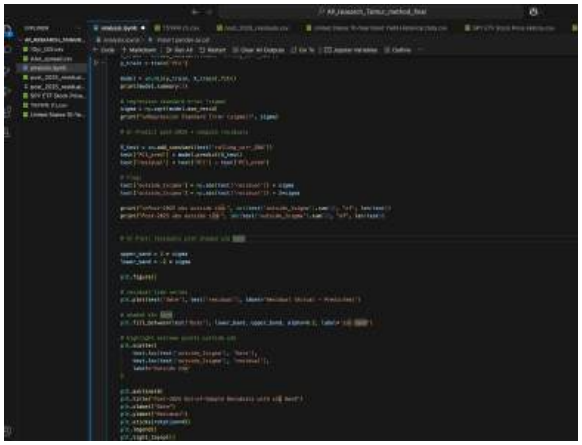
I'm happy to hear that you are interested in our VoxEU column.

To give you a quick answer to your question, I believe that the stock-bond correlation is in itself not sufficient for the insurance component because it is backward-looking. The idea of using the inflation risk premium is that it is more forward-looking. Moreover, there are other factors that determine the convenience yield, such as the supply of convenient assets (as discussed, for example, here: https://matthiasgne.github.io/website/DangerousLiaisons_December2025.pdf).

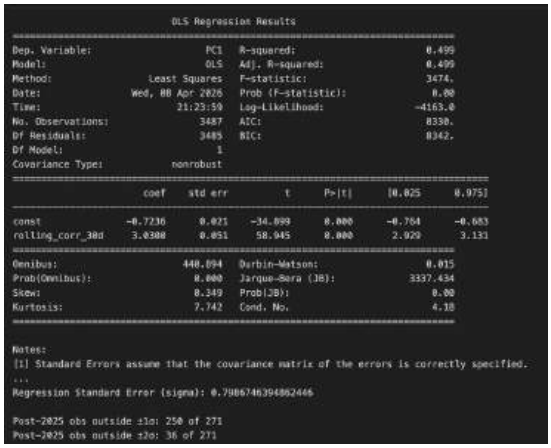
I hope this helps. Best of success with your research!

Best regards,
Matthias

Appendix F: Screenshot of Python code on Visual Code, with data sets on the left.



Appendix G: First regression results sheet, with R^2 value included.



OLS Regression Results

Dep. Variable:	PC1	R-squared:	0.499
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	3474.
Date:	Wed, 08 Apr 2026	Prob (F-statistic):	0.00
Time:	21:23:59	Log-Likelihood:	-4163.0
No. Observations:	3487	AIC:	8330.
Df Residuals:	3485	BIC:	8342.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7236	0.021	-34.899	0.000	-0.764	-0.683
rolling_corr_30d	3.0388	0.051	58.945	0.000	2.929	3.131

Omnibus:	440.094	Durbin-Watson:	0.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3337.434
Skew:	0.349	Prob(JB):	0.00
Kurtosis:	7.742	Cond. No.	4.18

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
...
Regression Standard Error (sigma): 0.7986746394862446

Post-2025 obs outside ±1σ: 250 of 271
Post-2025 obs outside ±2σ: 36 of 271

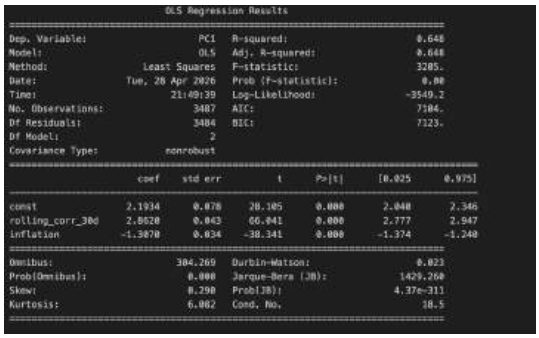
Appendix H: Actual residual data, illustrating where specifically *significant* negative residuals are.

post_2025_residuals

Date	residual	PC1	PC1_pred	below_2sigma
2025-01-02	-1.1406245854434500	-2.116294001321650	-0.9757694158782030	FALSE
2025-01-03	-1.1764935629544100	-2.139433418997540	-0.9629386560431350	FALSE
2025-01-06	-1.2182061996376900	-2.176905160308700	-0.9586989608710160	FALSE
2025-01-07	-1.2529873166989200	-2.217467599528520	-0.9644802828296030	FALSE
2025-01-08	-1.258083693693330	-2.228935525102010	-0.9706518314086850	FALSE
2025-01-10	-1.2479064739677500	-2.236985186321820	-0.989078712354065	FALSE
2025-01-13	-1.2706491785472000	-2.260129770660880	-0.9894805921136840	FALSE
2025-01-14	-1.290274800192950	-2.271794029434850	-0.9815192292419010	FALSE
2025-01-15	-1.2070389689886900	-2.247208113694560	-1.0401691447058700	FALSE
2025-01-16	-1.1965097218332800	-2.234792622095580	-1.0382829002623	FALSE
2025-01-17	-1.1943733657218900	-2.232346723750620	-1.037973358028730	FALSE
2025-01-21	-1.1456944866195500	-2.194181616241960	-1.0484871296224100	FALSE

Results illustrate that from the start of the regression residual analysis, there is underperformance; however, the negative residuals are not below 2 SD, therefore, these are not statistically significant deviations.

Appendix I: Second regression results sheet, with R^2 value included.



OLS Regression Results

Dep. Variable:	PC1	R-squared:	0.648
Model:	OLS	Adj. R-squared:	0.648
Method:	Least Squares	F-statistic:	3285.
Date:	Tue, 28 Apr 2026	Prob (F-statistic):	0.00
Time:	21:40:39	Log-Likelihood:	-3549.2
No. Observations:	3487	AIC:	7104.
Df Residuals:	3484	BIC:	7123.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1934	0.078	28.105	0.000	2.040	2.346
rolling_corr_30d	2.8620	0.043	66.041	0.000	2.777	2.947
inflation	-1.3070	0.034	-38.341	0.000	-1.374	-1.240

Omnibus:	304.269	Durbin-Watson:	0.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1429.260
Skew:	0.298	Prob(JB):	4.37e-311
Kurtosis:	6.002	Cond. No.	18.5

How Dance Enhances Memory: What Neuroscience Reveals About Movement, Music, and the Aging Brain By Zhirui Liu

Abstract

Most people think of dance as an art form or simply a way to stay physically active (Rehfeld et al.; Meng et al.). Neuroscience tells us a much more interesting story. Dancing requires the brain to coordinate movement, remember sequences, respond to music, process visual information, and interact with other people at the same time (Kandel et al.; Bear et al.). Because so many brain systems are working together during dance, researchers have started exploring whether dancing can strengthen memory and even help protect against age-related cognitive decline (Rehfeld et al.). This review examines findings from neuroscience and psychology to explain how dance affects memory formation (Erickson et al.; Rehfeld et al.), brain plasticity, and healthy aging. The evidence suggests that dance is much more than exercise. It is a complex brain activity that challenges multiple cognitive networks at once, making it one of the most engaging forms of physical activity for lifelong brain health.

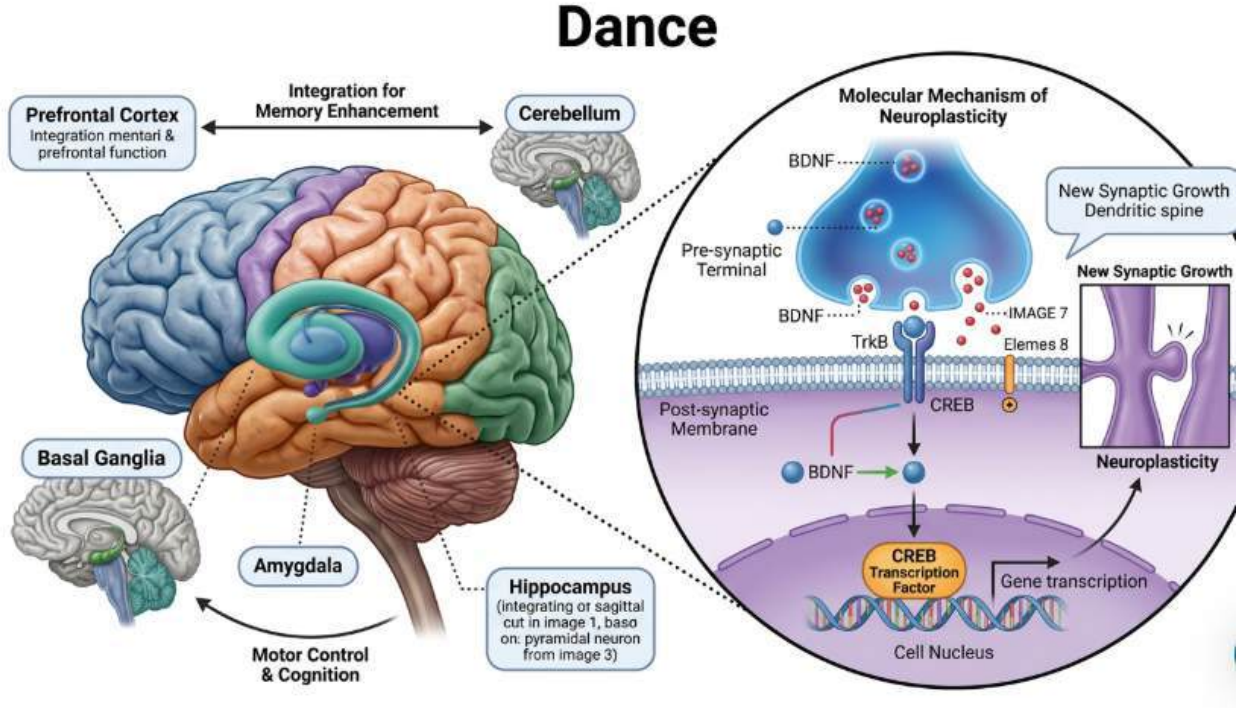
Introduction

Every movement we make begins inside the brain. Something as simple as taking a step requires communication between neurons, sensory systems, and muscles. Dancing makes this process much more demanding. A dancer must remember choreography, stay on rhythm, adjust body position, and often respond to partners or changing formations. All of these actions happen within seconds. Scientists have become increasingly interested in dance because it combines several factors already known to support brain health. Physical exercise increases blood flow(Cotman and Berchtold). Learning choreography strengthens memory. Music activates emotional centers. Social interaction reduces stress (Livingston et al.). Few activities combine all of these elements as naturally as dance. This raises an important question. Could dance be one of the most effective ways to improve memory throughout life?

Neural Systems Involved in Dance

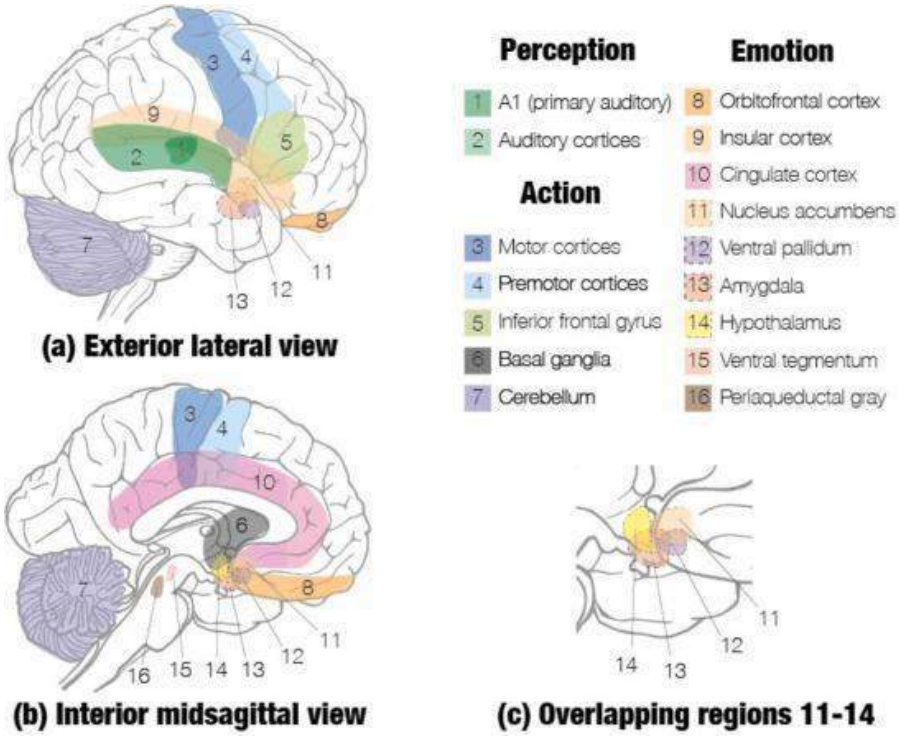
Memory is not stored in a single location. Instead, it depends on communication between many different brain regions. The hippocampus (Squire and Kandel) plays a central role in forming new memories (Squire). It helps people learn new information and remember experiences. The prefrontal cortex (Bear et al.) supports attention, planning, and working memory. The cerebellum (Kandel et al.) fine tunes movement and helps develop motor memory through repeated practice. The basal ganglia contribute to learning movement patterns until they become automatic. When someone learns a dance routine, these regions work together. The hippocampus remembers the order of movements. The cerebellum improves coordination after repeated practice. The prefrontal cortex focuses attention on timing and technique. As practice continues, the brain strengthens connections between neurons through a process called neuroplasticity. Neuroplasticity(Pascual-Leone et al.) allows the brain to reorganize itself by

forming new neural pathways. This ability exists throughout life, although it becomes less efficient with age. Activities that repeatedly challenge the brain encourage these pathways to become stronger. Dance provides exactly this kind of challenge.



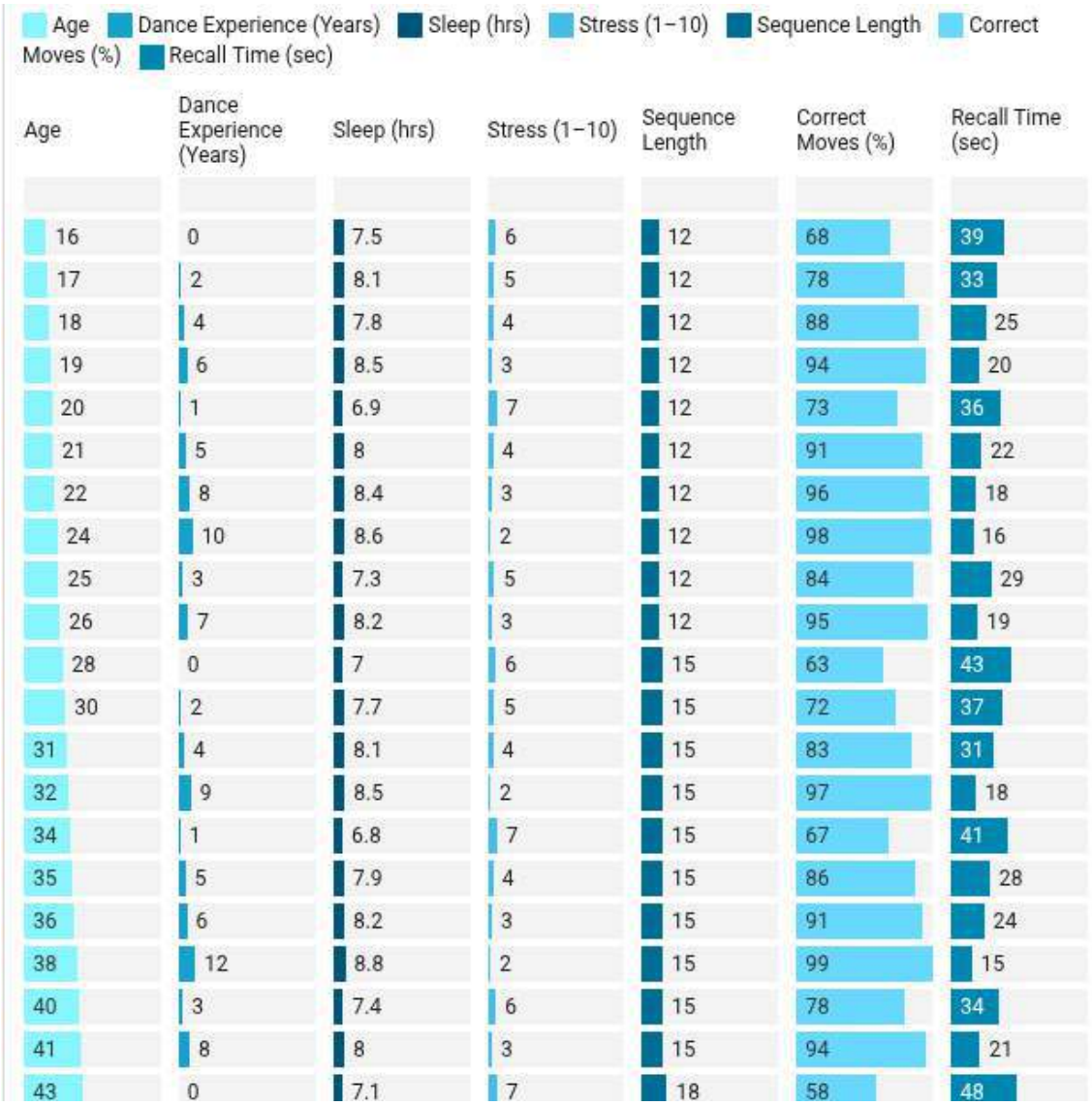
Dance as a Dual-Task Cognitive Activity

Many forms of exercise improve cardiovascular health, but dance asks much more of the brain. Running on a treadmill eventually becomes repetitive. Cycling often follows a predictable rhythm. Dance rarely stays the same. Dancers constantly learn new steps, adjust their posture, synchronize with music, and respond to mistakes without stopping the performance. Researchers sometimes describe dance as a "dual-task"(Rehfeld et al.) activity because it combines physical movement with continuous cognitive processing. For example, imagine learning a new dance routine. A dancer must remember the next movement while listening to music, maintaining balance, watching teammates, counting beats, and expressing emotion. These demands activate widespread brain networks instead of relying on only one skill. Because so many systems work together, dance may produce greater cognitive benefits than exercise alone.

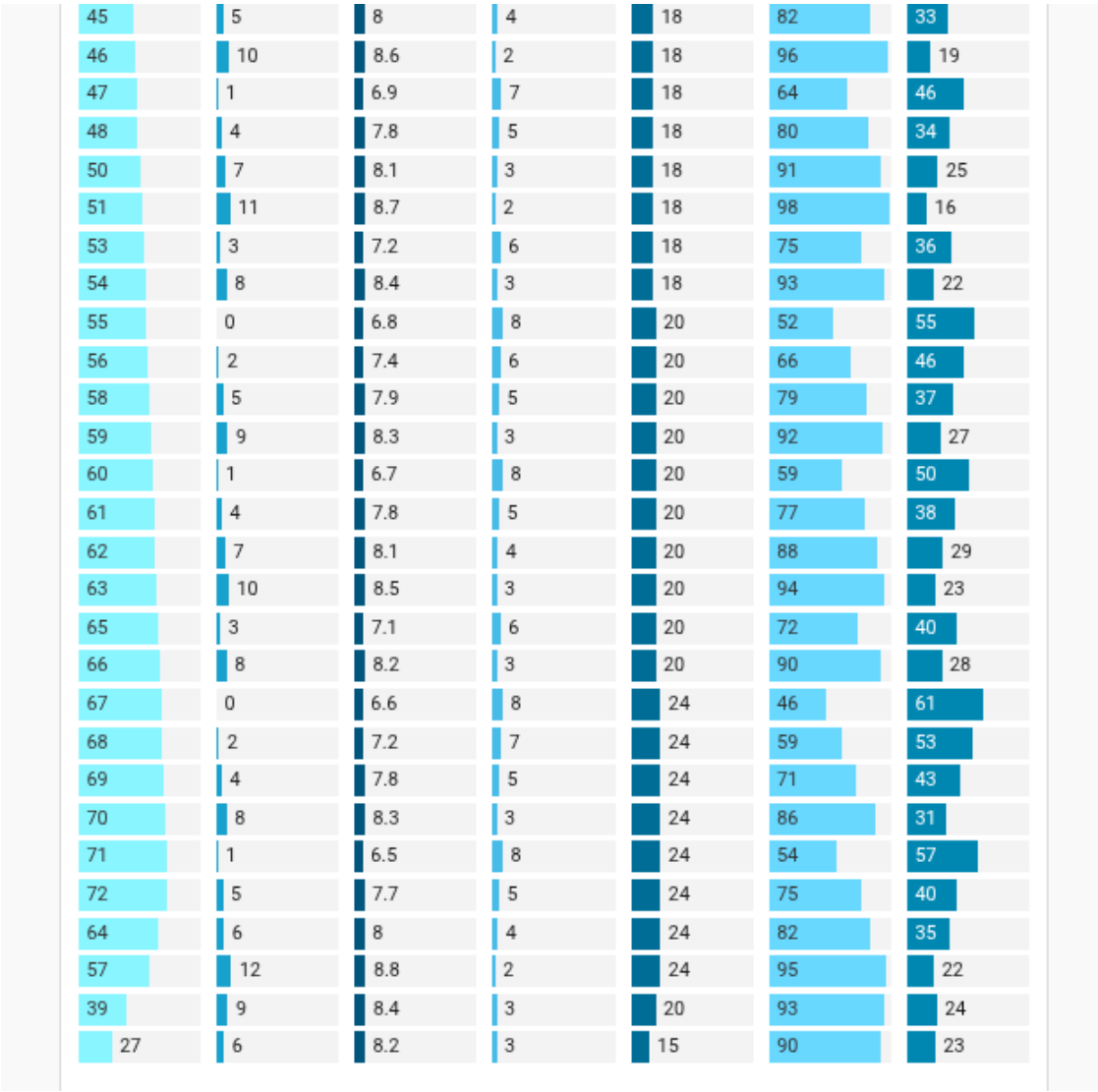


Neuroplasticity, BDNF, and Memory Formation

One of the brain's greatest strengths is its ability to change. Every time a person practices choreography, neurons that fire together become more strongly connected. This follows a basic principle in neuroscience often summarized as "cells that fire together wire together." Repeated movement also increases production of brain-derived neurotrophic factor, commonly called BDNF (Cotman and Berchtold). This protein supports neuron growth, strengthens synapses, and helps the brain form long-term memories. One cellular mechanism underlying memory formation is long-term potentiation (LTP), a persistent strengthening of synaptic connections following repeated activation. Increased BDNF expression during exercise helps support LTP, making learning and memory more efficient. Physical activity also increases oxygen delivery and blood circulation to the brain. Better blood flow provides neurons with more nutrients while helping remove waste products that accumulate over time. Together, these biological changes create an environment where learning becomes easier and memories become stronger.
Dance induces neuroplasticity through the upregulation of brain-derived neurotrophic factor (BDNF). While repetitive exercise increases BDNF, the constant novelty of choreography in dance triggers the BDNF-TrkB signaling pathway more intensely. When BDNF binds to the tropomyosin receptor kinase B (TrkB) receptor, it initiates the CREB (cyclic AMP response element-binding) protein cascade. This cascade activates genes responsible for synaptic remodeling and long-term potentiation (LTP), making dance a superior stimulator of structural brain changes compared to monotonous activities like walking or cycling.



Age	Dance Experience (Years)	Sleep (hrs)	Stress (1–10)	Sequence Length	Correct Moves (%)	Recall Time (sec)
16	0	7.5	6	12	68	39
17	2	8.1	5	12	78	33
18	4	7.8	4	12	88	25
19	6	8.5	3	12	94	20
20	1	6.9	7	12	73	36
21	5	8	4	12	91	22
22	8	8.4	3	12	96	18
24	10	8.6	2	12	98	16
25	3	7.3	5	12	84	29
26	7	8.2	3	12	95	19
28	0	7	6	15	63	43
30	2	7.7	5	15	72	37
31	4	8.1	4	15	83	31
32	9	8.5	2	15	97	18
34	1	6.8	7	15	67	41
35	5	7.9	4	15	86	28
36	6	8.2	3	15	91	24
38	12	8.8	2	15	99	15
40	3	7.4	6	15	78	34
41	8	8	3	15	94	21
43	0	7.1	7	18	58	48



45	5	8	4	18	82	33
46	10	8.6	2	18	96	19
47	1	6.9	7	18	64	46
48	4	7.8	5	18	80	34
50	7	8.1	3	18	91	25
51	11	8.7	2	18	98	16
53	3	7.2	6	18	75	36
54	8	8.4	3	18	93	22
55	0	6.8	8	20	52	55
56	2	7.4	6	20	66	46
58	5	7.9	5	20	79	37
59	9	8.3	3	20	92	27
60	1	6.7	8	20	59	50
61	4	7.8	5	20	77	38
62	7	8.1	4	20	88	29
63	10	8.5	3	20	94	23
65	3	7.1	6	20	72	40
66	8	8.2	3	20	90	28
67	0	6.6	8	24	46	61
68	2	7.2	7	24	59	53
69	4	7.8	5	24	71	43
70	8	8.3	3	24	86	31
71	1	6.5	8	24	54	57
72	5	7.7	5	24	75	40
64	6	8	4	24	82	35
57	12	8.8	2	24	95	22
39	9	8.4	3	20	93	24
27	6	8.2	3	15	90	23

Music, Emotion, and Memory

Dance rarely exists without music. Music activates several brain regions associated with emotion, reward, and memory. Emotional experiences are often remembered more clearly than neutral ones because the amygdala strengthens memory formation during emotionally meaningful events. When movement is paired with music, the brain receives information through multiple sensory pathways at once. Visual cues, sound, rhythm, and body movement all reinforce one another. This may explain why many dancers can remember routines learned years earlier. The choreography is connected not only to movement but also to music, emotions, and personal experiences. Instead of relying on a single memory pathway, dance creates several.

Social Interaction

Many dance styles involve working with other people. Whether performing in a group or dancing with a partner, participants must communicate without words. They watch facial

expressions, anticipate movements, and adjust their timing together. Social engagement itself has been linked to better cognitive health. Studies suggest that loneliness and social isolation may increase the risk of cognitive decline in older adults. Dance naturally encourages communication, teamwork, and emotional connection while simultaneously providing physical exercise. This combination may offer unique benefits that individual exercise cannot fully replicate.

Dance and Aging Brain

As people age, normal changes occur in the brain—memory becomes less efficient, processing speed slows, and learning new information requires more effort. Research suggests that dance takes advantage of the brain's lifelong ability to adapt by combining physical activity with continuous mental engagement. In a randomized controlled trial by Kattenstroth et al. (2013), older adults who participated in a six-month dance intervention demonstrated significant improvements in balance, reaction time, and cognitive flexibility compared with control participants. These tasks repeatedly activate the hippocampus, prefrontal cortex, and cerebellum, which are all important for learning and memory. Beyond the biological effects of improved blood flow and BDNF release, dancing often happens in groups, creating meaningful social interaction. While dance is not a cure for diseases such as Alzheimer's, structured dance programs may help older adults preserve cognitive function, maintain memory, and improve overall quality of life as they age.

Limitations and Future directions

Although many studies report positive effects of dance, sample sizes are often small and intervention lengths vary widely, making comparisons difficult. Additionally, some observed benefits may result from general physical exercise or social interaction rather than dance itself. Future research should aim to determine which dance styles—such as ballet, hip-hop, ballroom, or contemporary—provide the greatest cognitive benefit. Larger randomized controlled trials and advances in brain imaging will be necessary to observe how different styles influence neural networks over time. Understanding these relationships could help healthcare professionals develop affordable, enjoyable programs that support lifelong brain health.

Conclusion

Dance challenges the brain in ways that few other activities can. It combines movement, music, memory, emotion, and social interaction into a single experience. Neuroscience research increasingly suggests that these combined demands strengthen the brain's ability to learn and adapt. While many questions remain, the current evidence points toward an encouraging conclusion. Dance is not simply entertainment or exercise. It is a powerful cognitive activity that may support memory throughout life and help older adults remain mentally engaged as they age. As our understanding of the brain continues to grow, dance may become more than a performing art. It may become an important tool for promoting lifelong brain health.

Works Cited

Blumen, H. M. (2023). Noninvasive interventions in healthy and at-risk older adults: the case for social dancing. Innovation in Aging, 7(Suppl 1), 775. https://doi.org/10.1093/geroni/igad104.1263

Huang, C. S., Yan, Y. J., Luo, Y. T., Lin, R., & Li, H. (2023). Effects of dance therapy on cognitive and mental health in adults aged 55 years and older with mild cognitive impairment: a systematic review and meta-analysis. BMC Geriatrics, 23(1), 693. https://doi.org/10.1186/s12877-023-04406-y

Müller, P., Rehfeld, K., Schmicker, M., Hökelmann, A., & Dordevic, M. (2017). Evolution of neuroplasticity in response to physical activity in old age: the case for dancing. Frontiers in Aging Neuroscience, 9, 56. https://doi.org/10.3389/FNAGI.2017.00056

Rehfeld, K., Lüders, A., Hökelmann, A., Lessmann, V., & Kaufmann, J. (2018). Dance training is superior to repetitive physical exercise in inducing brain plasticity in the elderly. PLOS ONE, 13(7), e0196636. https://doi.org/10.1371/JOURNAL.PONE.0196636

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience, 14(2), 257-262. https://doi.org/10.1038/nn.2726

Tao, D. H., Awan-Scully, R., Ash, G. I., Pei, Z., & Gu, Y. (2023). The effectiveness of dance movement interventions for older adults with mild cognitive impairment, Alzheimer's disease, and dementia: a systematic scoping review and meta-analysis. Ageing Research Reviews, 91, 102120. https://doi.org/10.1016/j.arr.2023.102120

Zhu, Y., Zhong, Q., Ji, J., Ma, J., Wu, H., & Gu, S. (2022). Effects of dance interventions on cognitive function in older adults: A meta-analysis. Frontiers in Psychology, 13, 866665. https://pubmed.ncbi.nlm.nih.gov/39923448/

The Economic Effects of Emerging E-Commerce Platforms on U.S. Retail Competition
By Trevor Wu

Abstract

This paper investigates the economic effects of emerging service platforms, such as Temu, Shopee, Alibaba, and Shein, on smaller U.S. businesses and dominant companies like Amazon and eBay. Using strategic frameworks including SWOT analysis, PEST analysis, and Porter's Five Forces, this paper unpacks how these low-cost, globally focused platforms will and have reshaped the competitive landscape through aggressive digital marketing, efficient supply chains, and price disruption. While the established U.S. companies maintain advantages in infrastructure and brand trust, they now face increasing pressure from foreign entrants that exploit macroeconomic trends like inflation and shifting consumer preferences. The research finds that success in modern e-commerce depends on a firm's ability to balance cost leadership with innovation, respond to external threats, and adapt to changing market conditions, challenges that affect both small domestic sellers and large multinational companies.

1 Introduction

The global retail landscape is undergoing a great transformation as e-commerce continues to outpace traditional offline retail. Online platforms have reshaped consumer expectations for price, convenience, and delivery speed, forcing retailers to adapt or risk obsolescence. This shift has intensified competition not only between digital and brick-and-mortar businesses but also within the e-commerce sector itself, as new entrants and innovative business models disrupt established market leaders.

A key front in this competition is the growing rivalry between emerging non-U.S. e-commerce platforms and mature U.S. market leaders. Companies such as Temu, Shein, and Shopee, often based in rapidly expanding Asian markets, have leveraged low-cost manufacturing, aggressive digital marketing, and cross-border logistics to quickly capture global market share. Their rapid ascent poses a direct challenge to U.S. platforms like Amazon and eBay, which have long maintained domestic dominance through scale, infrastructure, and consumer trust. Analyzing how these companies compete provides insight into the shifting balance of power in global e-commerce.

This research applies established strategic frameworks to evaluate how U.S. e-commerce leaders are responding to rising foreign competitors, and how these dynamics are redefining the broader retail industry. To analyze this shifting competitive landscape, this paper draws on three key business analysis frameworks: SWOT analysis, PEST analysis, and Porter's Five Forces. SWOT analysis evaluates a company's internal strengths and weaknesses, along with external opportunities and threats, offering a balanced view of its competitive position. PEST analysis focuses on macro-environmental factors— political, economic, social, and technological—that influence a firm's strategy, especially when operating across international borders. Porter's Five Forces provides a structural view of industry competition by examining rivalry among existing

competitors, the threat of new entrants, supplier and buyer power, and substitute products or services. Together, these frameworks offer a comprehensive lens through which to understand how e-commerce platforms operate, compete, and adapt in an increasingly globalized and volatile market.

2 Background

2.1 Amazon

Founded in 1994, Amazon has evolved from an online bookstore into one of the most dominant e-commerce platforms in the world. Its success is largely built on a combination of cost leadership, technological innovation, and logistical scale. Amazon's cost leadership strategy enables it to offer competitive prices by leveraging economies of scale, global supplier networks, and tight control over overhead costs. Its technological innovation is evident in its investments in artificial intelligence, recommendation algorithms, cloud computing through AWS, and cashierless retail systems like Amazon Go, tools that enhance both customer experience and operational efficiency. Meanwhile, Amazon's logistical scale, anchored by hundreds of fulfillment centers worldwide, allows it to process and deliver millions of orders with speed and precision.Through services like Fulfillment by Amazon (FBA), which enables third-party sellers to store, pack, and ship their products using Amazon's logistics network, the company has not only expanded its product offerings but also raised consumer expectations for delivery speed, reliability, and convenience, an industry shift often referred to as the "Amazon Effect." In the U.S. and other developed markets, Amazon offers an extensive catalog of products across nearly every category and has expanded into adjacent industries such as cloud computing (AWS), video streaming (Prime Video), and cashier-less retail (Amazon Go)[17]. Its infrastructure-heavy model and Prime membership ecosystem have given it a near unmatched ability to control the end-to-end customer experience, reinforcing consumer loyalty and seller dependence.

2.2 eBay

eBay, by contrast, operates a fundamentally different model. As one of the earliest pioneers in online peer-to-peer commerce, eBay connects individual buyers and sellers without holding any inventory or managing logistics. Its marketplace is known for auctions, secondhand goods, collectibles, and hard-to-find items[18]. While it lacks Amazon's direct control over shipping and fulfillment, eBay benefits from a flexible, asset-light approach that allows for seller autonomy and a wide range of product diversity. Over the years, eBay has introduced innovations such as authenticity guarantees, global shipping programs, and AI-driven search tools to improve trust and usability[19]. Although eBay does not compete with Amazon on speed or operational scale, it retains a strong niche appeal, particularly among hobbyists, resellers, and collectors who prioritize value, rarity, or specialized product categories.

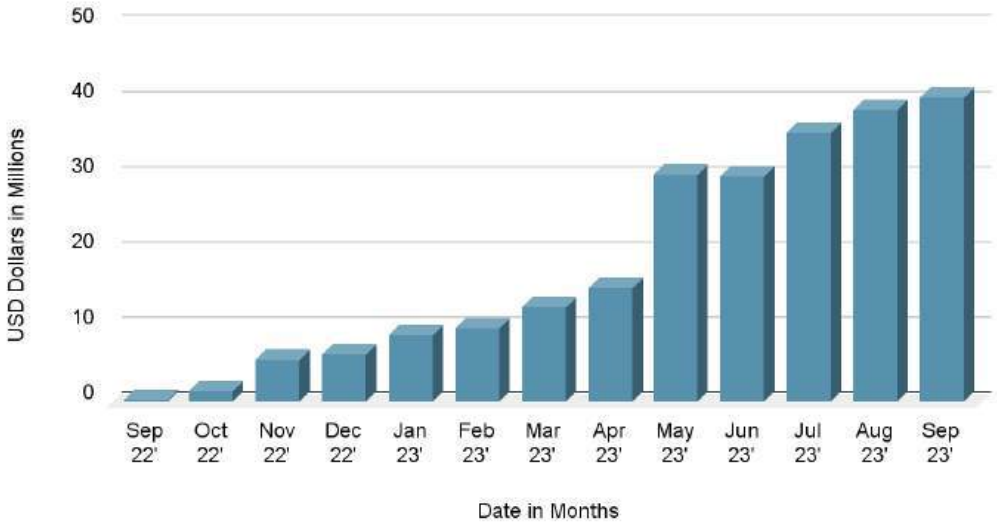
[17] ResearchGate 2023
[18] New York Times Archive 1999
[19] Reuters 2023

2.3 E-commerce

On the other side of the global market, a new generation of non-U.S. platforms, such as Shein, Temu, and Shopee, is rising quickly by targeting price-sensitive consumers and exploiting global supply chain efficiencies. These platforms typically operate with cost leadership strategies, outsourcing production to low-cost regions and minimizing overhead through streamlined logistics and digital-only storefronts[20]. Emerging non-U.S. platforms like Shein (founded in 2008), Shopee (2015), and Temu (2022) have rapidly gained global traction by targeting price-sensitive consumers through ultra-low-cost production, digital-only storefronts, and aggressive social media marketing. Backed by strong parent companies and lean logistics, these firms are challenging U.S. incumbents by offering trend-driven products at unbeatable prices, especially during inflationary periods. Shein, for instance, has disrupted the fashion industry with its ultra-fast production cycles and low prices, primarily targeting Gen Z consumers through viral social media marketing[21]. Temu, backed by Chinese tech giant PDD Holdings, uses a similar model but offers a wider variety of products beyond fashion, positioning itself as a discount superstore. Shopee, operated by Sea Group and dominant in Southeast Asia, combines aggressive regional expansion with localized payment systems and seller incentives[22]. While these platforms lack Amazon's infrastructure or eBay's niche credibility, they are gaining ground quickly, especially during inflationary periods, by appealing to global consumers seeking low prices and trend-driven products. This rapid rise is particularly evident in Temu's explosive global adoption. As shown in Figure 1, downloads of the Temu shopping app surged from near zero in late 2022 to over 40 million monthly downloads by September 2023. The sharp increase illustrates Temu's ability to scale quickly across global markets, leveraging aggressive marketing campaigns, viral promotions, and an expanding product variety to capture consumer attention.

[20] CNBC 2023
[21] Digital Commerce 360 2023
[22] Reuters 2019

Figure 1 - Global Downloads of Shopping APP Temu 2022-2023



Notes : Source App Economy Insights. 2025. The graph shows the rapid global adoption of the Temu shopping app from September 2022 to September 2023. Downloads started at nearly zero in late 2022, grew steadily through early 2023, and then spiked sharply in May 2023 to around 30 million. After this surge, growth continued through the summer, reaching over 40 million downloads by September 2023. This trend highlights Temu's explosive rise within just one year, driven by aggressive marketing, low-cost offerings, and its ability to quickly capture consumer attention worldwide.

3 Literature Review

One of the most widely used strategic planning tools in business analysis is the SWOT analysis, an acronym for strengths, weaknesses, opportunities, and threats. This framework allows organizations to systematically identify and evaluate both internal and external factors that may impact their success, guiding informed strategic decisions. Internal factors, strengths and weaknesses, are those over which the company has direct control, such as its resources, management practices, and organizational culture. External factors, opportunities and threats, stem from the market and macro-environment, including competitors, consumers, and regulatory changes. SWOT is particularly useful because it combines qualitative insights with quantitative metrics, giving firms an overview of their competitive position. For example, Amazon has used SWOT to understand how its highly efficient logistics network (a major strength) can cancel out rising labor costs, which is a weakness, while also going after opportunities in the quickly growing computing sector and diminishing threats from increasing regulatory inspection and questioning in the U.S. and Europe. By analyzing all four factors, companies can design or establish strategies that play to their advantages while addressing weaknesses as well as external challenges.

3.1 SWOT Analysis

The strengths component of SWOT focuses on what an organization does exceptionally well and the resources it can utilize to maintain its competitive edge. Strengths often include strong brand reputation, proprietary technology, large-scale operations, customer loyalty, or strategic partnerships. For instance, Apple's estimated global brand value of $516 billion in 2023 underscores its immense consumer trust and market presence, allowing it to maintain premium pricing even in competitive areas. Walmart's strength lies in its massive economies of scale: operating over 10,500 stores worldwide and leveraging a sophisticated supply chain, it keeps costs low and serves price-sensitive consumers effectively.[23] Moreover, Amazon's advanced fulfillment network enables next-day or same-day delivery for millions of items, a logistical strength unmatched by most competitors. Identifying such strengths allows companies to allocate resources strategically, ensuring these advantages are maximized in ways that enhance profitability and market share.

In contrast, weaknesses in SWOT are internal shortcomings or areas where the organization underperforms relative to peers. These vulnerabilities may stem from high debt loads, outdated technology, inconsistent branding, lack of profitability, or rigid organizational structures that limit flexibility. A notable example is Sears, which filed for bankruptcy in 2018 after decades of declining sales, a weakness driven by failure to adapt to digital retail and poor inventory management.[24] More recently, Wish.com serves as a cautionary tale: the platform's reputation suffered severely due to unreliable shipping times and poor product quality, leading to plummeting user trust and a 90% stock decline since its 2020 IPO.[25] Weaknesses can often amplify external threats if left unaddressed, which is why identifying and mitigating them is a critical part of strategic planning. Acknowledging weaknesses candidly allows companies to improve operations or pivot away from areas where they cannot realistically compete.

On the external side of SWOT, opportunities represent favorable conditions or emerging trends that an organization can exploit to grow revenue, enter new markets, or improve its competitive position. Opportunities often arise from demographic shifts, technological breakthroughs, regulatory changes, or unmet customer needs. For example, Shopify capitalized on the pandemic-era surge in online retail by providing scalable e-commerce solutions to small businesses, growing its merchant base to over 1.7 million by 2022.[26] Temu seized the opportunity of high inflation in the U.S., positioning itself as a low-cost alternative for increasingly price-conscious consumers. Another opportunity comes from geographic expansion; Shein, for example, expanded from China to become one of the most downloaded apps globally, reflecting the potential of tapping into underserved international markets.[27] Recognizing opportunities is crucial for sustaining growth and maintaining relevance in dynamic industries.

On the other hand, threats are external forces that could harm an organization's performance, profitability, or even survival if not addressed. Common threats include intense

[23] Brand Finance, 2023
[24] Peterson, 2018
[25] MarketWatch, 2023
[26] Shopify, 2022
[27] Sensor Tower, 2023

competition, regulatory crackdowns, macroeconomic instability, technological disruption, and supply chain shocks. Inflation, which peaked at over 9% in the U.S. in June 2022, posed a significant threat to many retailers, as it squeezed consumer purchasing power and increased operating costs.[28] This environment favored ultra-low-cost competitors like Shein and Temu, exacerbating the threat for mid-market and premium brands. Regulatory actions, such as proposed antitrust measures against large tech firms in the U.S. and EU, also illustrate how external policies can constrain even the largest players.[29] Companies that fail to anticipate and adapt to such threats risk rapid decline in market share or reputational damage.

3.2 PEST Analysis

The PEST framework complements SWOT by focusing exclusively on external macro-environmental factors, political, economic, social, and technological, that can affect a business. Unlike SWOT, which blends both internal and external analysis, PEST provides a much more structured perspective through which companies can evaluate the external forces shaping their operating environment, especially in international contexts where these factors may vary dramatically.[30] For example, as Temu expanded into the U.S., it had to navigate a different set of labor laws, consumer expectations, and tax structures compared to its home market in China. PEST enables organizations to anticipate risks and adapt to trends before they become existential challenges.

Political factors within PEST reflect how government policies, trade restrictions, tax regimes, labor laws, and political stability impact businesses. For instance, the U.S. and China trade war imposed tariffs of up to 25% on certain imports, increasing costs for platforms like Alibaba and forcing supply chain reevaluation.[31] California's Assembly Bill 5 (AB5), which reclassified gig workers as employees, increased operational costs for platforms dependent on flexible, low-cost labor models like Uber and DoorDash. Political stability is also vital: in countries experiencing political unrest, businesses face higher risks of disruption, making strategic planning more difficult.[32] Companies operating globally must therefore stay abreast of both domestic and foreign political developments.

Economic factors analyze metrics such as GDP growth, inflation, exchange rates, and interest rates, all of which influence consumer demand and operational costs. The strengthening of the U.S. dollar in 2022 made imported goods cheaper for American consumers, benefiting platforms like Temu, but hurting U.S. producers who could no longer compete effectively on price. Conversely, inflation reduced consumers' real purchasing power, shifting demand toward value-oriented retailers and away from premium brands.[33] Companies attuned to these shifts can

[28] BLS, 2022
[29] FTC, 2023
[30] General synthesis
[31] Reuters, 2019
[32] NYT, 2020
[33] WSJ, 2022

adjust pricing, marketing, and sourcing strategies to maintain competitiveness in changing economic climates.

Social factors encompass demographic shifts, cultural preferences, lifestyle changes, and public sentiment. Today's younger consumers value convenience, affordability, and digital-first experiences, fueling the rise of platforms like Shein and Temu.[34] At the same time, there is growing concern over sustainability and ethical sourcing, as seen in consumer backlash against fast fashion's environmental impact. Brands such as Patagonia have successfully differentiated themselves by aligning with these values, showing that social awareness can be both a threat and an opportunity depending on how companies respond.[35]

Technological factors refer to innovation, infrastructure, and research and development(R&D) investment, which increasingly determine market leadership. Amazon invested $73 billion in R&D in 2022, creating innovations like cashierless stores, drone delivery systems, and highly automated warehouses. These advancements enhance customer experience and reduce costs, setting a high bar for competitors.[36] Platforms that fail to innovate risk obsolescence as consumer expectations evolve. Technology also enables disruptive entrants to challenge incumbents by leveraging new tools, as seen with the rapid adoption of AI in personalized marketing and supply chain optimization.

3.3 Five Forces Structure

Michael Porter's Five Forces framework provides another perspective by analyzing industry structure and the intensity of competition. It examines competitive rivalry, threat of new entrants, supplier power, buyer power, and threat of substitutes to assess market attractiveness and profitability. This model helps companies understand where their industry stands and which forces to prioritize in strategic planning.

Competitive rivalry is intense in e-commerce, with Amazon, Walmart, Target, Temu, and Shein vying for market share. In 2023, Shein became the most downloaded shopping app in the U.S., highlighting the ferocity of competition and the pressure it places on incumbents to innovate or cut costs.[37] Companies operating in high-rivalry industries often face price wars and margin compression.

The threat of new entrants depends largely on entry barriers. While Amazon's scale and infrastructure create high barriers, platforms like Temu and Shein have demonstrated how digital marketing and low-cost sourcing can lower these barriers and allow rapid market entry. Bargaining power of suppliers is often low in e-commerce, as suppliers are fragmented and dependent on platforms for distribution.[38] Amazon and Walmart leverage their scale to negotiate favorable terms, further squeezing supplier margins. The bargaining power of buyers is high due to transparency and choice. Customers can compare prices instantly and switch providers easily,

[34] BLS, 2022
[35] UM, 2023
[36] CNBC, 2023
[37] Sensor Tower, 2023
[38] Lohr, 2022

pressuring firms to maintain low prices or offer superior value. The threat of substitutes remains a constant challenge, as consumers can switch to alternative products or brands that meet the same needs at lower prices. Fast-fashion and discount platforms exemplify how substitutes can displace incumbents quickly.

The cost leadership strategy is a deliberate approach to achieve the lowest operational costs in an industry. Walmart exemplifies this through its massive supply chain and aggressive cost control.[39] Temu and Shein follow a similar strategy by outsourcing production to low-cost regions and minimizing infrastructure investments. This allows them to offer products at lower prices, capturing price-sensitive market segments. However, such strategies come with thin margins and can leave companies extremely vulnerable to supply chain disruptions or increasing input costs.

The issue of anchoring prices to USD versus not anchored reflects how international platforms manage exchange rate risks. Platforms that price in USD provide U.S. consumers with predictable costs but expose foreign sellers to fluctuations in their home currencies.[40] Conversely, pricing in local currencies can create volatility that makes planning and profitability more difficult.

Understanding the macroeconomic context, including exchange rates, trade balances, and tariffs, is essential for international business. For example, during the U.S.-China trade war, tariffs raised import costs, hurting platforms like Alibaba but benefiting domestic competitors. Meanwhile, the strong dollar in 2022 made imports cheaper, intensifying competition from foreign platforms.[41] At the microeconomic level, firms adopt strategies based on cost leadership or product differentiation. Temu and Shein compete by offering ultra-low prices, while Apple and Etsy differentiate through premium quality, brand identity, and unique customer experiences.[42] Each approach appeals to different market segments, and both have advantages and drawbacks.[43] Tariffs protect domestic industries by raising the cost of imports, but also raise consumer prices. The U.S.-China tariffs were estimated to cost American consumers $1.4 billion monthly in higher prices, showing the trade-off between protectionism and affordability. Exchange rates influence competitive dynamics.[44] A strong dollar benefits importers but challenges domestic international exporters. Large platforms like Amazon and Alibaba hedge against currency risks through global diversification.

Finally, R&D investment is key to maintaining a competitive edge. Amazon's substantial R&D budget has resulted in innovations that smaller competitors cannot match, demonstrating how continuous investment in technology can create long-term advantages.[45] When comparing business strategies, cost leadership dominates platforms like Temu and Shopee, while product

[39] Lohr, 2022
[40] WSJ, 2022
[41] NY Fed, 2019
[42] Shopify, 2022
[43] Brand Finance, 2023
[44] WSJ, 2022
[45] CNBC, 2023

differentiation characterizes brands like Apple and Etsy. Both strategies can succeed but require alignment with market conditions and customer expectations.

4 Business Analysis

Amazon's business model truly conveys the power of scale and hybrid retail in e-commerce. Shipping over 500 million items annually, Amazon combines cost leadership and differentiation strategies by operating the company both as a direct retailer to consumers and as a third-party marketplace. This hybrid approach allows Amazon to offer consumers a broad selection at competitive prices while also ensuring small businesses who rely on its Fulfillment by Amazon (FBA) services. Amazon's FBA program boosts sellers' reach and sales by giving them access to Prime shipping and Amazon's logistics network, while offering consumers faster, more reliable delivery and raising competitive barriers in the e-commerce market. In contrast to single purpose models like Temu or Shopee, Amazon occupies both the role of competitor and platform provider, enabling it to take value from multiple points in the supply chain. The company's ability to fuse scale with market economics reveals its positioning with the cost leadership strategies described in the literature review, while also differentiating itself through proprietary services like Amazon Web Services and cashierless stores.

4.1 SWOT Analysis Applied

A SWOT analysis of Amazon highlights the internal and external factors shaping its competitiveness. Among the many of its strengths, Amazon's highly advanced logistics infrastructure enables next and same-day delivery across the U.S., reinforcing its position of value[46]. Additionally, its trusted brand and Amazon Web Services(AWS) cloud services provide diversified revenue streams and technological dominance. However, weaknesses remain, including ongoing labor disputes, antitrust investigation, and low retail margins, which erode profitability and potentially expose reputational risks. On the opportunity side, Amazon continues to innovate in physical retail through Whole Foods and Amazon Go, while pursuing international expansion into emerging markets. Meanwhile, threats from regulatory actions, rising labor costs, and increasing competition from low-cost rivals like Temu and Shein threaten its market dominance. These dynamics reflect the importance of recognizing how internal weaknesses can magnify external threats if left unaddressed—a key insight from the review.

4.2 Five Forces Applied

Applying Porter's Five Forces reveals Amazon's strategic positioning in a highly competitive industry. Competitive rivalry is extremely intense, with incumbents like Walmart and new firms such as Temu exerting pressure. However, Amazon mitigates this through its scale and proprietary services, setting the benchmark known as the Amazon Effect. Entry barriers and expectations remain high, given Amazon's capital-intensive logistics and data-driven personalization capabilities. Supplier power is low due to a broad base of vendors reliant on its

[46] IvyPanda, 2023

platform, while buyer power is high because of price transparency and low switching costs. Finally, substitutes like unique marketplaces or brick-and-mortar options remain marginal threats, illustrating how Amazon's innovation sustains its advantage even in a market characterized by high buyer power and rivalry.

4.3 PEST Applied

Looking through the lens of the PEST framework, Amazon's external environment further contextualizes its challenges and opportunities. Politically, antitrust litigation and regulatory scrutiny or questioning in the U.S. and EU expose the company to significant legal risks[47]. In markets like India, foreign ownership regulations limit how Amazon manages sellers, and in Australia and Canada, evolving tax and consumer laws raise compliance costs. These political pressures directly impact Amazon's ability to replicate its U.S. model globally. Economically, inflation and evolving consumer price sensitivity force Amazon to sustain low prices, further pressuring margins. Currency fluctuations also matter: a strong U.S. dollar benefits import pricing but hurts U.S. exports. Meanwhile, global supply chain disruptions, like those during COVID-19, have driven Amazon to diversify suppliers and invest in regional fulfillment networks to mitigate risk. Socially, consumers demand convenience and sustainability, which Amazon partially meets through digital-first offerings but struggles with in ethical sourcing. Mobile commerce growth has made app optimization essential, while sustainability concerns push Amazon toward carbon-neutral delivery goals by 2040, despite criticism over emissions from its logistics network. Cultural adaptation also plays a role—Prime benefits in markets like Japan and India include localized perks tailored to regional preferences.Technologically, Amazon remains a leader, investing $73 billion in R&D in 2022[48], creating innovations such as drone delivery, AI-driven personalization, and cashierless stores. These investments reflect the necessity of continuous technological development to adapt to shifting consumer expectations and regulatory landscapes.

Apple's business strategy contrasts with Amazon's strategy, showcasing the strength of product differentiation and expertise over a variety of fields. In a SWOT analysis, Apple's strengths include its globally recognized brand (valued at $516 billion in 2023)[49], its integrated ecosystem, and consistent innovation in hardware and services. Some weaknesses include high reliance on iPhone revenues, vulnerability to supply chain disruptions, and criticism over labor practices in its supplier system. Opportunities lie in expanding health-related technologies, AR/VR, and subscription-based services. Some threats could include aggressive Android competitors, economic downturns impacting premium product demand, and global antitrust actions. Porter's model puts Apple in a high-rivalry environment, at the same time it leverages brand loyalty and ecosystem lock-in to mitigate buyer power and maintain high margins. Entry barriers remain significant due to Apple's proprietary technologies and economies of scope. The PEST analysis highlights challenges from U.S.-China trade tensions, a strong dollar affecting

[47] FTC, 2023
[48] R&D World, 2022
[49] WIPO, 2023

exports, and rising consumer expectations around sustainability. Apple's R&D spending of $27.7 billion in 2023 supports innovation in chips, wearables, and augmented reality, exemplifying the long-term advantage of continuous R&D investment aligned with the literature review's emphasis on innovation as a competitive differentiator.

Temu represents a new kind of cost leadership competitor, disrupting U.S. e-commerce by exploiting low-cost production and aggressive digital marketing. Other platforms like Shopee dominate in Southeast Asia, leveraging localized marketing, while Shein leads in fast fashion through ultra-fast production cycles. Alibaba competes globally through wholesale (Alibaba.com) and retail (AliExpress), focusing on scale and merchant variety. These competitors highlight that Amazon's moat is not insurmountable—price disruption, cultural targeting, and alternative supply chains are credible threats. In a SWOT analysis, Temu's strengths lie in its ultra-low prices, viral marketing campaigns, and cost-efficient, direct-from-China sourcing[50]. However, weaknesses include poor quality perception, inconsistent delivery times, and limited customer service capabilities. Opportunities include expanding into more durable goods and leveraging data analytics to improve personalization. Threats arise from regulatory scrutiny over data privacy and tariffs stemming from U.S.-China trade tensions. Porter's Five Forces highlight that Temu intensifies rivalry by lowering price floors and benefiting from lower entry barriers through a digital-only model. However, its lack of infrastructure and brand trust make it vulnerable to substitutes and consumer skepticism. A PEST analysis illustrates these vulnerabilities further: politically, U.S.-China trade conflicts and potential tariffs pose serious risks[51]; economically, inflation helps Temu attract price-conscious shoppers but limits profitability; socially, Temu appeals to a younger demographic yet faces backlash over sustainability and labor practices; technologically, it lags behind competitors in proprietary innovation, relying instead on scale and price.

4.4 Ebay Contrast

Amazon and eBay are two of the most recognized online marketplaces in the world, but their business models, competitive strategies, and operational structures diverge in significant ways. Both platforms connect millions of buyers and sellers globally, benefit from network effects, and operate in multi-billion-dollar markets. However, the similarities largely end there, as Amazon's hybrid approach combines direct retail with a third-party managed marketplace, while eBay functions purely as an intermediary between buyers and sellers without owning inventory or handling fulfillment at scale. From a SWOT perspective, Amazon's strengths include its vertically integrated logistics network, proprietary technology stack, and brand trust—factors that allow it to enforce strict quality control and ensure rapid delivery through services like Fulfillment by Amazon (FBA). eBay's strengths, by contrast, lie in its flexible platform model, which supports a broader diversity of product categories, including niche and second-hand goods that Amazon often overlooks. For example, eBay dominates in collectibles,

[50] CNBC, 2023
[51] Reuters, 2019

vintage apparel, and automotive parts, categories where authenticity, rarity, and buyer-seller negotiation matter more than two-day shipping. However, eBay's lack of direct control over fulfillment can be a weakness in terms of delivery speed and customer experience consistency—areas where Amazon's FBA system excels. Both face threats from emerging global competitors like Temu, Shopee, and Shein, as well as shifting consumer preferences toward specialized marketplaces. Both platforms Amazon and Ebay benefit from network effects, but Amazon's vertically integrated ecosystem ensures tighter control over the buyer experience, whereas eBay relies on seller autonomy and community-driven trust systems.

These analyses demonstrate the distinct strategies of three major players and their alignment with the frameworks detailed in the literature review. Amazon exemplifies a hybrid of cost leadership and differentiation, supported by continuous R&D investment and high barriers to entry. Ebay represents a leaner, more specialized marketplace model, proving that a niche focus can compete and coexist alongside large market companies.Temu represents disruptive cost leadership, exploiting macroeconomic conditions and digital marketing to gain a foothold, albeit at the risk of sustainability and regulatory challenges. Together, these cases illustrate the trade-offs between cost leadership and differentiation, the importance of innovation, and the influence of macroeconomic and political forces—all principles highlighted in your theoretical discussion.

5 Conclusion

The rise of global e-commerce competitors like Temu, Shein, and Shopee signals a major shift in the industry, challenging the long-standing dominance of U.S. platforms such as Amazon and eBay. These new foreign companies have disrupted the market through low-cost strategies, digital marketing, and supply chain efficiency, especially during times of economic uncertainty.

Strategic frameworks like SWOT, PEST, and Porter's Five Forces reveal that while Amazon maintains strengths in scale, logistics, and innovation, it faces growing pressure from regulatory scrutiny, shifting consumer expectations, and intense competition. Meanwhile, companies like eBay retain a unique and niche appeal but must adapt to stay competitive.

Going forward, success in global e-commerce will depend on companies' ability to balance cost leadership with innovation, respond to external pressures, and evolve with consumer values. Those that fail to adapt to new situations and expectations risk losing relevance in an increasingly fast-paced and unpredictable market.

Works Cited

Apple Inc. Form 10-K, Annual Report 2023. Cupertino, CA: Apple Inc., 2023. https://www.apple.com/newsroom/pdfs/FY23_Annual_Report.pdf.

App Economy Insights. "Pinduoduo: Overtaking Alibaba." App Economy Insights (Substack), December 5, 2023. Accessed August 29, 2025. https://www.appeconomyinsights.com/p/pinduoduo-overtaking-alibaba

Axios. “Amazon Illegally Maintains a Monopoly, FTC Alleges in Landmark Lawsuit.” September 26, 2023. https://www.axios.com/2023/09/26/ftc-amazon-monopoly-lawsuit.

Bureau of Labor Statistics. “Consumer Price Index — June 2022.” July 13, 2022. https://www.bls.gov/news.release/archives/cpi_07132022.htm.

Business Insider. “Sears Is Closing 46 More Stores — Here’s the List of Closing Locations.” August 23, 2018. https://www.businessinsider.com/sears-kmart-close-46-stores-list-2018-8.

CNBC. “Temu’s Rise in the U.S. Market and Challenges.” November 15, 2023. https://www.cnbc.com/2023/11/15/temu-challenges-and-opportunities-in-us.html.

Digital Commerce 360. “Shein’s Lead under Fire as Temu Tops U.S. App Charts.” March 10, 2023. https://www.digitalcommerce360.com/2023/03/10/sheins-lead-under-fire-chinese-owned-app-temu-tops-us-charts/.

Federal Trade Commission. “FTC Sues Amazon for Illegally Maintaining Monopoly Power.” September 26, 2023. https://www.ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power.

IvyPanda. “The Amazon Company’s Organizational Analysis.” 2023. https://ivypanda.com/essays/the-amazon-companys-organizational-analysis.

MarketWatch. “Wish Parent ContextLogic’s Stock Sinks to Penny-Stock Territory as Company’s Troubles Mount.” January 3, 2023. https://www.marketwatch.com/story/wish-parent-contextlogics-stock-sinks-to-penny-stock-territory-as-companys-troubles-mount-11674653417.

Michigan Journal of Economics. “Op-Ed: Patagonia Proves the Success of Sustainable Corporations.” November 22, 2023. https://sites.lsa.umich.edu/mje/2023/11/22/op-ed-patagonia-proves-the-success-of-sustainable-corporations/.

New York Federal Reserve. “Who’s Paying for the Trade War? Evidence from the U.S. Tariffs.” May 23, 2019. https://libertystreeteconomics.newyorkfed.org/2019/05/whos-paying-for-the-trade-war-evidence-from-the-us-tariffs/.

New York Times (Archive). “A Business and Technology Book Review; Selling the Internet, Making Money.” October 10, 1999.

https://archive.nytimes.com/www.nytimes.com/library/review/101099ecommerce-review.html.

R&D World. “How Much Does Amazon Invest in R&D? Here’s an Estimate.” July 12, 2022. https://www.rdworldonline.com/how-much-does-amazon-invest-in-rd-heres-an-estimate/.

ResearchGate. “Amazon: An Organizational Review of an E-Commerce Company.” 2023. https://www.researchgate.net/publication/367453219_AMAZON_An_organizational_review_of_an_e-commerce_company.

Reuters. “Apple Faces Antitrust Scrutiny and Trade Tension Risks.” June 15, 2023. https://www.reuters.com/technology/apple-us-china-trade-tensions-2023-06-15/.

Reuters. “Timeline: Key Dates in the U.S.-China Trade War.” May 10, 2019. https://www.reuters.com/article/business/timeline-key-dates-in-the-us-china-trade-war-idUSKCN1SE2OZ.

Shopify. Shopify 2022 Annual Report. March 1, 2023. https://s27.q4cdn.com/572064924/files/doc_financials/2022/ar/Form-40-F-2022.pdf.

WIPO. “Global 500 2023: The World’s Most Valuable Brands.” January 2023. https://brandfinance.com/reports/global-500-2023.

Wall Street Journal. “The Surging U.S. Dollar Creates Winners and Losers in the Global Economy.” October 8, 2022. https://www.wsj.com/articles/the-surging-u-s-dollar-creates-winners-and-losers-in-the-global-economy-11665161977.

The Anatomy of Dismissal: Ancient Bodies, Ancient Control
Gender Discrimination in Women's Medical Treatments Across the Classical World
By Manuela Mejía-Fanjul

Abstract

Through a full-scale analysis of medical textbooks, treatments, and anecdotal evidence across Mesopotamia, Ancient Egypt, Ancient Greece, and Ancient Rome, this paper examines the historical roots of gender discrimination embedded within medical theory and practice. First, by contrasting the approaches to female anatomy in sources such as the Kahun Gynecological Papyrus, Hippocratic writings, and Galenic texts, the study measures both the differences and similarities. Additionally, by comparing the medical traditions across these societies, the analysis reveals a persistent, cross-cultural tendency to overlook female pain, suggesting that ancient medical theories have influenced modern clinical practices for millennia. Ultimately, this work aims to contribute to the broader historical discussion concerning the use of medicine as a mechanism of power and societal control.

Key Words "Ancient Medicine", "Women's Health", "Gender Discrimination", "Gynecology", "Classic Antiquity", "Hippocratic Corpus", "Social Control"

Introduction

Understanding the historical treatment of women's health requires situating ancient medical practices within broader societal structures. Across Mesopotamia, Egypt, Greece, and Rome, medical texts were not merely neutral descriptions of the human body; they functioned as cultural instruments that reinforced social hierarchies, gender norms, and moral expectations. The examination of female bodies through the lens of ancient medical knowledge reveals a persistent undervaluing of women's experiences and a systemic marginalization of their pain. By tracing the historical roots of this discrimination, we gain insight not only into the social positioning of women in antiquity but also into the enduring ways these attitudes continue to shape contemporary medicine and societal perceptions of female bodies.

Mesopotamia

Women's health and bodies were treated as agents of reproduction and fertility, rather than human beings with voices, pains, and emotions. In other words, women were treated based on just how much they could give to society through their offspring. Because of this, the majority of records relating to women's health in the Mesopotamian era focus on birth, fertility, and menstrual difficulties.

Some of the most common female ailments of the empire include infertility, constant menstruation, miscarriages, childbirth, and irregular/concerning vaginal discharge. Though valid ailments, the treatments were less reputable and effective. For example, to induce birth, doctors had pregnant female patients drink grape and cumin while walking around their hometown (Iraq

Museum, Baghdad, no. 153). However, women's pain did not just end there. After the induction of labor, the birthing process was, for many, excruciating and at times even fatal. The Mesopotamian doctors' approach to managing labor and birthing pain was the utilization of chants. Although the perspective on the actual intentions of the chants, whether to save the fetus or the mother, is mixed, there is evidence of a popular chant that was said to be repeated around the time of a birth:

> The [birthing] mother is surrounded by the dust[s] of death, Like a chariot, she is surrounded by the dust[s] of battle, Like a plough, she is surrounded by the dust[s] of woods, Like a fighting warrior, she is fighting in her blood. Weakened, her eyes don't see, her lips are covered, she doesn't open [them], the fate of death and the fate of silence, her eyes [...] (De La Fuente, "La medicina Mesopotámica," 1615).

Some historians characterize these cantations as ancient spells, but regardless of their form, the passage represents the sheer danger women faced in the act of childbirth. In fact, the passage mentions the fate of death for the women giving birth. Although there aren't statistics and formalized records that have been found by archeologists, by using artifacts representative of the era's cultures, historians have been able to recognize the intense harm these women suffered during birth. The fact that the Mesopotamians had chants describing the death of women in childbirth proves how common it was, showing the pain and perishing that was in store for brand-new moms.

Generally, besides childbirth, female health complaints were relatively comparable to today's health concerns. A common complaint from females in Mesopotamia was lactation difficulties. Many women complained of pain, discomfort, or lactation scarcity. Despite the widespread nature of these complaints, nothing was done by medical practitioners to solve or even research the issue.

In terms of menstrual cycles, the Babylonians, in particular, were focused and concerned with vaginal blood flow and their lack of control over the cyclical nature of it. A medical record states: "nu tar-su, lā ipparrasū," which directly translates to "it cannot be stopped" (De La Fuente, "La medicina mesopotámica," 1615). Historians question the tone in which this was meant, whether the physicians at the time wrote it in a scholarly fashion or a panicked cry for help. Regardless, there was significant worry surrounding acute blood loss, particularly in young women with heavy menstrual cycles (De La Fuente, "La medicina mesopotámica," 1615).

Similarly, there are records of women's vaginas shedding copious amounts of fluids. The fluids themselves are unclear; whether it could be menstrual blood, regular discharge, urine, or sexual fluids isn't recorded, but their remedies and causes are. Cuneiform reading, "diš munus ú.ḫi.a ze-ru-te šu-ku-ul a.meš ina šà gal4.la-šá ma-gal du-ku" translates to "If a woman has been given herbs of hate-(magic) to eat (and because of this) 'fluids' flow abundantly from her vagina" (De La Fuente, "La medicina Mesopotámica," 1615). Speculating that the reasoning

behind the abundance of fluid flowing from the vagina is related to magic or witchcraft further divides the space between women's care and facts, giving way to millennia of discrimination.

Also contributing to the discrimination is the concept that women's pain should not be seen. Again in cuneiform, a record states: "diš munus mú-šá tar-mí mú-šá nu [innamir], this text translated directly to "If a woman's blood is cut off so that her blood is not [seen]" (De La Fuente, "La medicina Mesopotámica," 1615). Although not the typical Modern English sentence structure, this text provides historical context on a significant piece of the medical discrimination puzzle: visibility. The first section of the clause is confusing and borderline illegible; however, the recovered word [seen] gives way to a larger conclusion, that in Mesopotamia, women's pain was supposed to be invisible. This invisibility was most likely to maintain an image of purity for the male population, as copious amounts of blood were deeply associated with pregnancy loss and, therefore, past sexual experiences.

Miscarriages were particularly common in Mesopotamia because of the lack of research and awareness around the issue. However, there was some knowledge recorded on the matter in cuneiform: "diš munus šá 3 iti múd-šú du-ak ana šà šà-šà [nu šub]" (De La Fuente, "La medicina mesopotámica," 1615). Translated to English, it reads, "If a woman who is three months (pregnant) bleeds, (remedy) so that she does [not abort] her fetus". The ancient Mesopotamians had a relatively good understanding of the signs of miscarriages, and the mention of "remedy" suggests that they had some idea of how to prevent them in general. Besides these specific quotes, the Mesopotamians kept extensive records of their health; however, one significant issue of clarity remains. The word "Mû" referred to several different bodily fluids that are all signs of different organ systems or ailments, but the specific fluid depended on the context of the ailment. From blood to water to amniotic fluid, essentially any document with that word mentioned remains unclear, leaving a vast gap in the knowledge of women's health.

Ancient Egypt

Similarly, Ancient Egypt's stories are deeply ingrained in modern medicine. Buried in the history of a modern-day beautiful, bustling, urban tourist attraction lies the rich and intricate stories of Ancient Egyptian women and their bodies. Unlike some other ancient civilizations mentioned in this text or omitted, Egyptian women were treated remarkably, enjoying significant legal, economic, and social rights, which varied depending on their region and social class. From a medical perspective, women's healthcare was mainly based on tricks that aren't necessarily rooted in fact. For example, remedies that we would call "old wives' tales" gained popularity as the empire aged, from garlic cloves placed in the vaginal canal to determine fertility to tying ropes around a woman's neck to gauge gestational age. These tricks, centered around myths, formed the core of the Ancient Egyptian medical system (Al-Shamma, "Miscarriages in Ancient Mesopotamia: Etiology, Remedies and Incantations," 492-516).

Although the tricks ranged from different disorders and ailments, many centered around menstruation and contributed to the image of a cycle in general. Beliefs around menstrual blood protecting from evil and inducing pregnancy if rubbed on one's thighs spread rapidly throughout

ancient culture. Although the relationship between menstruation and equality for Ancient Egyptians was anything but ordinary, the ideology that menstrual blood protected people was a large catalyst in the movement of period equity in Ancient Egypt. Over time, women would have assets that other civilizations would see as luxuries, such as separate rooms for menstruating privately and having their husbands stay home from work so the menstruator could rest. Women in Ancient Egypt were treated remarkably well during their cycle.

The causes of these pleasures were plentiful. Although the thought that menstrual blood could protect from evil was definitely a factor in the treatment of periods, it was more complicated than that. Menstruation was frequently associated with cleansing and purification. "Hesmen", the Old Egyptian word for menstruation, was synonymous with the meaning "purification" (Diamond, "The Blood of Isis: Menstruation in Ancient Egypt").

Interestingly, Ancient Egyptian menstrual technology was pioneering for its time, from softened papyrus leaves to absorb menstrual blood to papyrus records of the oldest known texts on gynecological health (Strochlic, "From Shame to Celebration: The Long History of Periods"). The oldest known text about gynecology was titled *The Papyrus Kahun*. Several additional mythological tricks were recorded in the text,

> "If you want to know if a woman will give birth, you should place emmer (wheat) and barley in two bags. Let her urinate on them daily. If they both sprout, she will give birth. If only the barley sprouts, it will be a boy. If only the emmer sprouts, it will be a girl. If neither sprouts, she will not give birth" (Karamanou, "The Theory of the Four Humours").

Although not completely accurate, the Ancient Egyptians began developing clues to pregnancy tests. In modern-day medicine, urine is an accurate determinant of pregnancy, though barley and emmer are not key indicators of the sex of an unborn child (*The Ebers Papyrus*, trans. B. Ebell).

It is more important than ever to learn, understand, and continue to acknowledge the roots of the rights that many have had taken away from them in today's modern, contemporary world. Abortions and contraceptives are incredibly controversial in modern politics, and the roots of controversy and secrecy surrounding stigma existed long before modern procedures.

The Papyrus Kahun mentions contraception and provides instructions on how to prescribe it: "A prescription to prevent conception: ground acacia, dates, and honey. Place in soft fiber and insert into the vagina." Contraceptives seem to have less stigma when compared to abortions due to the sheer amount of texts on contraceptive measures" (Karamanou, "Four Humours").

In terms of abortions, the Egyptian Ebers Papyrus, a text similar to the Kahun, does actively mention different approaches to performing abortions. However, it remains unclear whether abortions actually happened and their legality. The methods appearing in a published text lead historians to believe that there weren't legal issues surrounding the termination of

pregnancies; however, the lack of recorded procedures does point to a societal stigma that prevented women from terminating their pregnancies (*The Ebers Papyrus*).

Importantly, women complaining of rheumatoid disorders were met with denial and annoyance. Physicians advised them to improve their complexion and attitude to be less disturbing to others in their pain (Al-Shamma, "Miscarriages in Ancient Mesopotamia: Etiology, Remedies and Incantations," 505). Although not representative of the rest of women's health in Ancient Egypt, the message that women's pain was disturbing to others is threaded throughout ancient and contemporary history, threading medical sexism through time and space.

Though Middle Eastern and African studies are vital, most of modern American culture is framed by European ideas, making Ancient European civilizations the framework for not only Europe itself, but the United States as well. In terms of sexism in the medical field, European ideas began in Ancient Greece.

Ancient Greece

Greek culture was heavily situated in philosophy, making philosophers vital members of society and builders of ideologies. Famous philosophers such as Hippocrates, Aristotle, and later Galen systematically downplayed women's health issues because of their interpretation of the female anatomy. They genuinely believed and reported that women's bodies were defective versions of men's bodies. Robert Garland, a present-day British philologist and historian, cites Aristotle as the main culprit for this issue, stating, "From Aristotle's perspective, women were failed males" (Garland, *Daily Life of the Ancient Greeks*). Though ideologies produced by them were certainly flawed, Greek philosophers created a standard for medicine unbeknownst to mankind at the time. For example, Hippocrates wrote *Magni Hippocratis*, a published work made up of around seventy individual books. The work was composed over the course of 80 years, from 450 to 350 BCE, in several European cities, and was later translated into dozens of languages.

Although the work is no longer whole or comprehensive, archives have begun to piece together the writings. For example, it is known that the writings included discussions of abortion- a topic that has now been proven to spread through ancient civilizations, outliving its controversy. Additionally, the writings included the Hippocratic oath, a pledge to do no harm that is routinely used today in medical school.

Ultimately, *Magni Hippocratis* was based and composed entirely on the concept of the Four Humors, a medical ideology Hippocratis himself created and firmly believed in. The theory first states that there are four fundamental fluids in the human body: blood, phlegm, yellow bile, and black bile. These four fluids or humors each corresponded to different substances and aspects of the natural world. Each is paired with a temperature, a moisture, an element, and a season. Blood was typically correlated with heat, wetness, air, and spring; phlegm was indicative of coldness, wetness, water, and winter. Similarly, yellow bile was paired with heat, dryness, fire, and summer; and, finally, black bile was correlated with coldness, dryness, earth, and autumn (Learning Mole, "The Four Humours").

Hippocrates believed that a balance of these four humors meant general health, but an imbalance led to "eucrasia," or a condition that required treatment. This led to three theories that evolved to explain the menstrual period.

Hippocrates himself proposed the first; he believed that a period was simply the excess of heat in the female body. This was rooted in his earlier belief that blood was heat. Ultimately, many individuals found this to lack explanation and substantiation. So, the Aristotelian Model was born. This model explained that periods existed for the sole purpose of purging the female body of impure substances. The final model is relatively accurate. Titled the Galen Model, it was developed later in history, but it presented the idea that a fetus feeds off menstrual blood, and the lack of a fetus led to a period (McNeill, "Abortion Is Ancient History, and That Matters Today"). Ultimately, these three theories battled for the role of the truth about menstruation; it is not clear which was widely adopted then.

Despite subtle accuracy spreading slowly, an inaccurate theory that would cause the course of medicine to quickly evolve as truth: the wandering womb. The philosopher Plato believed the womb is a wandering, wild creature that would get aggravated and offended if not impregnated. This theory somehow resonated with other physicians and led to the wandering womb's connection to breast cancer and hysteria (D'Ettorre, Bersi, and D'Ettorre, "History of Childbirth and Obstetrics in Ancient Rome," 196–198). Although the connection to breast cancer never quite stuck, hysteria did stick as one of the incorrect foundational theories of women's medicine that would come to discriminate and gaslight women for millennia to come.

The term hysteria stems from the Greek word hysterika, which translates directly to uterus. Originally, it was just a wandering womb, but over time, ancient physicians and philosophers began to hypothesize that it led women to experience unsettled, extreme emotions. Throughout the next several centuries, hysteria transformed from a gynecological diagnosis to a psychiatric one, all because of the characterization of women's uterus in Ancient Greece.

Ancient Rome

Learning from Greece, Ancient Rome, a bustling city of innovation, was actually one of the first civilizations to build on the civilizations before it in medicine. Roman medicine was heavily influenced by Greek practices, causing women's bodies to be seen and treated as functional instruments in reproduction.

To understand the context of women's health, it's important to first grasp the sheer danger that came with childbirth. A hazardous circumstance for both the mother and child, there were extremely high mortality rates as traditional early Roman medicine centered on folklore and religious practices. Over time, however, Hippocratic medicine became more popular, and with its rise came innovative forms of care. This marks a pivotal era of medicinal history, the transition from religious care to obstetrics (D'Ettorre, Bersi, and D'Ettorre, "History of Childbirth and Obstetrics in Ancient Rome," 197). With this shift came other ideologies, from vaginal suppositories to keep women more comfortable to genuine concern for women with chronic pain (Scarfo, *Hysteria in the Hippocratic Corpus*).

Soranus of Ephesus is sometimes viewed as having single-handedly changed the course of medical care forever by introducing the concept of compassion. As a physician, he began the movement to touch on female topics in medical training, giving new doctors a background of ailments. This education allowed women not to be met with ignorance when explaining their pain, a massive advancement. This certainly didn't come with a lack of discoveries on his end; from learning how to handle a baby in breach to birth control methods, it seems Soranus was constantly in search of new information to fuel his practices (King, "The Hippocratic Corpus and the History of Hysteria," 1-17).

He was also an established author; In *On Chronic Diseases*, Soranus and a colleague display compassion and flexibility in their care, an unheard-of phenomenon. Essentially, Soranus is responsible for the personalization of medicine, for the monitoring of symptoms to alter treatments, and kindness through medicine (King and Hall, "Contraception and Abortion in Ancient Egypt, Greece and Rome: From Papyrus to Philosophy," 332-338).

Although he had great accomplishments, there were some shortcomings. In his written work, On Midwifery and the Diseases of Women, he highlighted the importance of hydration while also mentioning the different aspects and factors when determining when a young woman's period will arrive. He cites temperatures, imagination, flavorful foods, and provocative photos as factors to catalyze a woman's menstruation (Kālidāsa, *The Birth of the War-God*). Of course, since then, these theories have been proven irrelevant to the timing of a young girl's first period, but it's important to recognize that, for the time, wondering and researching about female anatomy was a progressive stride in medicine.

Continuing in revolutionizing women's health, Soranus developed a birth control method that was definitely more reliable than the alternatives before him. His method was called the "coitus interruptus"; it involved removing the penis from the vagina promptly during intercourse to avoid fertilization. Shockingly, this method has since then stayed popular; today it is more commonly known as "the pull-out method" (Scarfo, *Hysteria in the Hippocratic Corpus*, 110). Besides this specific method, Ancient Roman birth control was complicated and unclear, as the line between contraception and abortion was extremely blurred.

Soranus, however, was open to the concept of abortion as a treatment plan. In fact, in his writings, he stated that abortions are acceptable:

> ...to prevent subsequent danger in parturition if the uterus is small and not capable of accommodating the complete development, or if the uterus at its orifice has knobby swellings and fissures, or if some similar difficulty is involved (Scarfo, *Hysteria in the Hippocratic Corpus*, 110).

Generally, it was societally acceptable to terminate a pregnancy for health reasons; however, depending on class, for emotional or quality of life reasons, it varied.

From what historians have pieced together, it is clear that the stigma around abortion was heavily reliant on whether the woman in question was enslaved or free. Although enslaved

women lacked respect in Ancient Rome, they held more respect for determining whether or not to terminate a pregnancy than free women did. Essentially, because enslaved women were seen as property, the upper Roman class typically viewed their motherhood as inconvenient for them. Therefore, it wasn't atypical to see an enslaved pregnant woman praised for deciding to have an abortion. However, it was common for freewomen to be condemned for the very same reasoning (Scarfo, *Hysteria in the Hippocratic Corpus*, 110).

Generally, ancient civilizations managed the issue of women's health inadequately. Dismissing symptoms, physicians concluded that because of wandering uteruses, excessive heat, and a female's unappealing complexion, women were failed males. As a modern-day reader, that may seem absurd to read; in fact, you are probably questioning how they could have possibly landed on those conclusions. The reality is that those conclusions continue to be the foundation of modern-day medicine and pharmaceuticals. The only change that has really happened between these ancient civilizations and today is that large corporations have managed to disguise their sexism as convenience or the woman's lack of knowledge.

Works Cited

Al-Shamma, Safaa M. “Miscarriages in Ancient Mesopotamia: Etiology, Remedies and Incantations.” Quarterly Journal of the College of Arts 80 (2017): 492–516. https://qarts.journals.ekb.eg/article_114473_15c001f11447915467a0698b46b89b03.pdf

D’Ettorre, M., P. Bersi, and L. D’Ettorre. “History of Childbirth and Obstetrics in Ancient Rome.” The Clinical Journal of Pain 23, no. 2 (February 2007): 196–198. https://pubmed.ncbi.nlm.nih.gov/17355293/

De La Fuente, J. G. L. “La medicina Mesopotámica.” Asclepio: Revista de Historia de la Medicina y de la Ciencia66, no. 2 (2014). Accessed December 9, 2025. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0211-95362014000200002

Garland, Robert. Daily Life of the Ancient Greeks. London: Greenwood Press, 1998.

Kālidāsa. The Birth of the War-God: A Poem by Kālidāsa. Translated by Ralph T. H. Griffith. 1879; Project Gutenberg, 2023. Accessed December 9, 2025. https://www.gutenberg.org/ebooks/72544

Karamanou, Marianna. “The Theory of the Four Humours.” Encyclopedia.pub. Accessed December 9, 2025. https://encyclopedia.pub/entry/30441

Learning Mole. “The Four Humours.” Learning Mole (blog). Accessed December 9, 2025. https://learningmole.com/the-four-humours/

The British Museum. “Ancient Healthcare Fit for a King.” British Museum Blog. Accessed December 9, 2025. https://www.britishmuseum.org/blog/ancient-healthcare-fit-king

Iraq Museum, Baghdad. Photograph of museum object no. 153. Photograph by Anmar Fadhil. n.d.

McNeill, Leila. “Abortion Is Ancient History, and That Matters Today.” CNN, June 23, 2023. https://www.cnn.com/2023/06/23/health/abortion-is-ancient-history-and-that-matters-today-scn

Scarfo, Barbara N. “Hysteria in the Hippocratic Corpus.” PhD diss., McMaster University, 2020. https://macsphere.mcmaster.ca/bitstream/11375/25397/2/Scarfo_Barbara_N_2020April_PhD.pdf

King, Helen. “The Hippocratic Corpus and the History of Hysteria.” The Classical Quarterly 63, no. 1 (2013): 1–17. https://pubmed.ncbi.nlm.nih.gov/24041856/

King, Helen, and Lesley A. Hall. “Contraception and Abortion in Ancient Egypt, Greece and Rome: From Papyrus to Philosophy.” The European Journal of Contraception and Reproductive Health Care 21, no. 4 (2016): 332–338. https://pubmed.ncbi.nlm.nih.gov/26946682/

Diamond, Kelly-Anne. “The Blood of Isis: Menstruation in Ancient Egypt.” Isiopolis (blog), November 23, 2014. https://isiopolis.com/2014/11/23/the-blood-of-isis/#:~:text=A%20young%20woman's%20first%20menstruation,her%20or%20him%20from%20evil

Strochlic, Nina. “From Shame to Celebration: The Long History of Periods.” National Geographic, November 17, 2021. https://www.nationalgeographic.com/history/article/periods-menstruation-women-history-ancient-egypt

The Ebers Papyrus. Translated by B. Ebell. New York: D. Appleton and Co., 1937.

Polarizability of PEG-600 through VQE By Keerthan Challa

Abstract

The polarizability of polyethylene-glycol-600 (PEG) was found by using modern Qiskit tools and the Variational Quantum Eigensolver (VQE). The previous time it was done by Dr. Murthy and Dr. Naidu in a lab setting in 1977, almost 50 years ago (Murthy and Naidu 551). Without access to the resources, I used an algorithm to determine the ground-state energies when fields were applied to PEG's monomer, ethylene oxide.

The mean molecular polarizability equation was used to determine ethylene oxide's polarizability with the ground states. Assuming additivity, the polarizability was then multiplied by the quotient between the polymer and ethylene oxide's molecular weight to scale to the polymer. A basis set is a collection of functions. Two different bases (6-31+G* and aug-cc-pVDZ) with varying degrees of accuracy were tested through the algorithm, with 6-31+G* being the smaller basis of the two. The results were then compared with the literature monomer polarizability ("Experimental Data for C2H4O") and Dr. Murthy and Dr. Naidu's polymer polarizability (Murthy and Naidu 552). The values were close to the baselines, with the most accurate calculation being within approximately 4% of Dr. Murthy and Dr. Naidu's baseline. This paper demonstrates a modern application of quantum computing to reproduce an experiment.

Introduction

Polarizability is the tendency of an electron cloud to be distorted by an external electric field. It can be applied to determine London-dispersion forces, molecular reactivity, and optical refraction. London-dispersion forces result from instantaneous dipoles, resulting in attraction between molecules. Polarizability determines the effectiveness of the attraction. Molecular reactivity describes how easily a molecule goes through chemical change. The ease of the shifts within the electron clouds can be represented with polarizability. Optical refraction describes light bending when passing through materials. When light interacts with the molecules, polarizability can be used to describe how strongly the electrons respond and oscillate. This study aims to determine the mean polarizability of polyethylene-glycol (PEG) using a Variational Quantum Eigensolver (VQE).

My grandfather, S. V. Naidu, along with Dr. Murthy, found the polarizability of PEG-600 using magnetooptic rotation and the Faraday effect (Murthy and Naidu 551). Magnetooptic rotation had been used on many occasions for simple molecules and applying it to polymers was a new practice (Murthy and Naidu 551).

To determine the polymer's polarizability, Naidu et al. measured the Verdet constant with a photocell and operational amplifier using benzene as a standard (Murthy and Naidu 551). The Verdet constant represents the rotation per unit of magnetic field strength for the material. Naidu et al. continued with measuring the optical refraction using a Pulfrich refractometer, giving the refractive index. Density was then measured with a pycnometer at 25°C. Finally, Dr. Naidu and

Dr. Murthy found the molecular weight and the number of electrons. Using a formula previously derived in a paper by Dr. Murthy and Dr. Rao, Dr. Naidu and Dr. Murthy determined the polarizability of PEG-600 to be 5.639×10^{-23} cm^3 (Murthy and Naidu 552).

Formula used to determine polarizability:
$\alpha M = [c'nMp\delta / \rho(n^2+2)^2]^{(1/2)}$
(mean polarizability) = √{ [(constant) × (refractive index) × (molecular weight) × (effective number of electrons) × (Verdet constant)] / [(density) × ((refractive index)² + 2)²] }

I came across my grandfather's paper around 3 months ago and have been researching to recreate his experiment using modern tools. Since I didn't have access to lab or equipment, I attempted to recreate his results with simulation. While doing research, I came across the foundational paper on VQE (Peruzzo et al.), which explained how the algorithm could be used to compute molecular ground-state energies. After exploring deeper, I came across IBM's Qiskit library, a Python library for creating quantum simulations ("VQE"). The documentation explained how to find ground states, conversions from atomic orbitals to molecular orbitals, optimizers, ansatz, mappers, and finally VQE ("PySCFDriver"; "ElectronicEnergy"; "Mappers"; "StatevectorEstimator"; "VQE"). After weeks of reading through documentation and building code, I had created a mean polarizability calculator using ground-state energies of the field-perturbed molecule.

In the experiment, two bases were tested with varying degrees of accuracy to algorithmically determine the mean polarizability of the polymer. The experiment is proof of concept for using VQE to determine polarizability of molecules. The paper discusses the methodology, limitations, and findings of the approach, comparing the results with existing baselines. VQE's accuracy will depend on basis size, with larger bases being more accurate and numerically stable.

Methodology

In order to prevent high demand for processing power, the experiment used PEG-600's monomer unit: ethylene oxide. The molecule was constructed using a STO-3G basis to describe its electrons and later tested with 6-31+G* and aug-cc-pVDZ ("PySCFDriver"). To confirm the molecule properly loaded, a Hartree-Fock method was used to get the reference ground-state energy. Møller-Plesset's second-order perturbation theory was used to account for energy missed by Hartree-Fock.

After the reference ground-state was set, the mean polarizability was determined through classical methods. The molecule was set up again with its coordinates, and the same three bases (STO-3G, 6-31+G*, aug-cc-pVDZ) were tested. Seven field configurations were set up, with a ±0.001 a.u. field applied along the x, y, and z coordinates. For each field, the dipole integrals were built and multiplied by the field to override the Hamiltonian core. Restricted Hartree-Fock was then applied to converge and yield the ground states values. Restricted Hartree-Fock forces

pairs of electrons with opposite spins to occupy the same orbital, which occurs in ethylene oxide due to its closed-shell structure. After collecting the 7 energies, the finite-difference polarizability formula was applied, and the final values were converted to cm^3.

Finite-Difference Polarizability Formula:
$\alpha = -[E(+F) + E(-F) - 2E(0)] / F^2$
(polarizability) = − [(energy at +field) + (energy at −field) − 2×(energy at zero field)] / (field strength)²

After setting up the baseline, the VQE solver was set up ("VQE"). The three bases (STO-3G, 6-31+G*, aug-cc-pVDZ) were tested again. To simulate orbitals, eight qubits were utilized to keep processing within the limits of my computer's capacity. Four spatial orbitals with two spin states each needed eight qubits to represent them. Other orbitals were kept in a frozen state. The orbitals in the active space were two occupied orbitals (orbitals with electrons in ground-state) and two virtual orbitals (orbitals empty in ground-state) around the Fermi level ("Hamiltonians"). A field of ±0.001 was applied just like before. The fields were applied before applying Restricted Hartree-Fock, so orbitals stay relaxed under the field. CASCI was then used to extract the 1 and 2 electron integrals from the active space ("pyscf.mcscf"; Sun et al., "A General Second Order").

The electronic energy was found by using those integrals and the nuclear repulsion energy property was added ("ElectronicEnergy"). The Bravyi-Kitaev mapper was used because it creates shorter Pauli strings than Jordan-Wigner mapper, resulting in increased efficiency ("Mappers"). The ansatz selected was Unitary Coupled Cluster (UCC) because it lets me choose which excitations should be included. In the experiment, both single and double excitations were included. The optimizer was the L-BFGS-B, a gradient-based method, chosen because it reliably converged better with larger basis sets compared to the gradient-free COBYLA optimizer ("VQE"). The energy expectation values were computed using a statevector-based estimator ("StatevectorEstimator").

1000 iterations of VQE were run with the optimizer. The final ground-state energy was determined by summing the VQE's result, the frozen-core electronic energy, and the nuclear repulsion energy. Since the nuclear repulsion energy was included in both the frozen-core electronic energy term and set separately, nuclear repulsion was subtracted once to adjust for the overlap. Then, the mean polarizability was recorded and calculated using the finite-difference polarizability formula. The polymer was assumed to have an additive property. The mean polarizability was multiplied by 13.6, the ratio between PEG-600's molecular weight and the weight of ethylene oxide to estimate the final polymer's polarizability.

The resulting polarizabilities were compared to the literature value ("Experimental Data for C2H4O") as well as the value derived from magnetooptic paper (Murthy and Naidu 551).

Results

The classical basis was compared to the VQE to cross-check if the values seemed accurate. STO-3G was also tested but excluded from the table. STO-3G's classical value had a monomer polarizability around 61% lower than the literature value ("Experimental Data for C2H4O"), and the results were unstable across various tested optimizers. The VQE values with the same basis were more unstable, giving results over 250 times the literature values. The monomer values were compared to the literature value of ethylene oxide ("Experimental Data for C2H4O"), while the scaled values were compared to Dr. Naidu and Dr. Murthy's baseline (Murthy and Naidu 552). The general trend illustrates that the larger basis tends to be more accurate. The method with the least percent error was the VQE CASCI with the aug-cc-pVDZ basis, with a literature monomer error of around ten percent, and a polymer error of around 4.1 percent.

Method	Basis	α (monomer, cm^3)	α (PEG-600 scaled, cm^3)	% Diff. vs. Lit.	% Diff. vs. M&N
Classical HF	6-31+G*	3.554×10^{-24}	4.83×10^{-23}	~20% low	~14% low
Classical HF	aug-cc-pVDZ	3.96×10^{-24}	5.23×10^{-23}	~11% low	~7% low
VQE	6-31+G*, CASCI (4e,4o)	3.617×10^{-24}	4.92×10^{-23}	~18% low	~13% low
VQE	aug-cc-pVDZ, CASCI (4e,4o)	3.978×10^{-24}	5.41×10^{-23}	~10% low	~4.1% low
Literature, monomer	—	~4.43×10^{-24}	—	reference	reference
Dr. Murthy and Dr. Naidu (1977), PEG-600	—	—	5.639×10^{-23}	reference	reference

Discussion

The maximum number of iterations controls the convergence of the VQE optimization loop ("VQE"). The maxiter sets the number of iterations. In each iteration, the optimizer adjusts the parameters to minimize the expected Hamiltonian value. If a maxiter is set too low, a ground-state value will be produced before reaching the energy minimum. In the experiment, a maxiter of 1000 was used for all bases, to minimize the risk of premature convergence.

One explanation on why the STO-3G basis with the VQE method produced underestimated mean polarizability is the lack of a diffuse basis function. Polarizability measures how easily an electron cloud deforms in response to an applied field. The deformation occurs in outer regions, away from the nucleus. Diffuse functions capture the deformations further from the nucleus, allowing the basis to capture the electron cloud response. Basis functions, such as STO-3G are too contracted to the nucleus, which affects their accuracy at computing the outer electron cloud deformation. The bases, aug-cc-pVDZ and 6-31+G*, include diffuse functions, allowing them to capture the outer distortions. Even a higher maxiter of 2000 didn't solve the instability and produced values over 280 times larger than the literature value ("Experimental

Data for C2H4O"). The results indicated that the issue was with the basis, and not the number of iterations.

I used a 4 electron, 4 orbital active space to keep within the processing limits of my computer. However, the monomer itself has 24 electrons and 19 atomic orbitals. To represent the full active space, significantly more qubits would be used with exponentially deeper circuits, infeasible on my laptop. The 4 electron, 4 orbital active space was used to balance accuracy and computational power ("pyscf.mcscf"). The Hilbert space scales exponentially as 2^n, where n is the number of orbitals. 2^(19*2) states was reduced to 2^(4*2) states, and significant computation was saved.

The polarizability of PEG-600 was calculated by summing the polarizabilities of the monomer units. However, assuming additivity treats the monomers as independent molecules, and ignores intermolecular forces. Specifically, additivity ignores the electron distortions occurring between each monomer unit in the polymer chain. While the values showed close agreement with both reference values, measuring the exact discrepancy between the values is difficult. The literature monomer value ("Experimental Data for C2H4O") and the value derived from the polymer and monomer's molecular weight differ by approximately 6.4%, suggesting that the additivity assumption creates a discrepancy independent of the VQE and Classical calculations.

Percent difference = |literature value − derived value| / literature value × 100
= $|4.43\times10^{-24} - 4.146\times10^{-24}| / 4.43\times10^{-24} \times 100$
= 6.4%

VQE outperformed the Hartree-Fock calculations in accuracy and successfully demonstrated that polarizability gets more accurate with larger basis sets. To reiterate, the purpose of this study wasn't to replace classical models, but to demonstrate that quantum algorithms can be applied to chemically relevant problems (Peruzzo et al.). The similarity of the VQE and Classical values suggests that VQE captured the electronic structure response correctly, even with the active-space and basis set restrictions.

The mean polarizability (isotropic polarizability) is calculated as the average of the three diagonal components of the polarizability tensor. For some molecules, the polarizability tensor has high variation between different axes. As a result, the mean polarizability obscures the directional differences. This means the mean scalar polarizability value is a simplification of the polarizability tensor, excluding direction.

If I did the experiment again, I would try to use larger active spaces to see how close to the literature value I could get with VQE ("Experimental Data for C2H4O"). However, using larger active spaces would require more qubits, deeper quantum circuits and harder optimization. I would also try using zero-noise extrapolation: a technique where the same circuit is run at multiple noise levels and then estimating the result when there is zero noise ("Understanding Quantum Errors"). If I do get access to the resources, I would try to utilize real quantum

hardware. However, that would also introduce some problems although it would be more realistic. For example, real quantum hardware introduces vulnerabilities, such as interactions with the environment, bit-flip errors, and gate imperfections ("Understanding Quantum Errors").

Conclusion

The paper explored the application of VQE to determine the polarizability of the PEG-600 polymer. By setting up ground-state baselines, testing multiple bases, and recording results, the most accurate value was found to be around 4% away from the Dr. Murthy and Dr. Naidu's measured polarizability (Murthy and Naidu 552). The results demonstrated that larger bases tend to be more accurate, with around a 9% accuracy difference between the tested bases using VQE. Despite using a completely different method from the one my grandfather used 50 years ago, the produced values landed close to his original measurements. In doing so, the study demonstrates that modern quantum computing like VQE can be applied to real scientific problems (Peruzzo et al.).

Works Cited

Murthy, V. Rama, and S. V. Naidu. "Magnetooptic Rotation and Molecular Polarizability of PEG." Journal of Quantitative Spectroscopy and Radiative Transfer, vol. 19, 1978, pp. 551-552.

Peruzzo, Alberto, et al. "A Variational Eigenvalue Solver on a Photonic Quantum Processor." Nature Communications, vol. 5, 2014, article no. 4213, doi:10.1038/ncomms5213.

"Experimental Data for C2H4O (Ethylene Oxide)." NIST Computational Chemistry Comparison and Benchmark Database, National Institute of Standards and Technology, cccbdb.nist.gov/exp2x.asp?casno=75218.

"Understanding Quantum Errors: Challenges and Solutions in Quantum Computing." SpinQ, 1 Aug. 2025, www.spinquanta.com/news-detail/understanding-quantum-errors-challenges-and-solutions-in-quantum-computing.

"Hamiltonians." Quantum Chemistry with VQE, IBM Quantum Learning, qiskit.qotlabs.org/learning/courses/quantum-chem-with-vqe/hamiltonian-construction.

"pyscf.mcscf." PySCF Documentation, pyscf.org/pyscf_api_docs/pyscf.mcscf.html.

Sun, Qiming, et al. "A General Second Order Complete Active Space Self-Consistent-Field Solver for Large-Scale Systems." Chemical Physics Letters, vol. 683, 2017, pp. 291-297.

Sun, Qiming, et al. "Recent Developments in the PySCF Program Package." The Journal of Chemical Physics, vol. 153, no. 2, 2020, article no. 024109, doi:10.1063/5.0006074.

"PySCFDriver." Qiskit Nature Documentation, IBM Quantum, qiskit-community.github.io/qiskit-nature/stubs/qiskit_nature.second_q.drivers.PySCFDriver.html.

"ElectronicEnergy." Qiskit Nature Documentation, IBM Quantum, qiskit-community.github.io/qiskit-nature/stubs/qiskit_nature.second_q.hamiltonians.ElectronicEnergy.html.

"Mappers (qiskit_nature.second_q.mappers)." Qiskit Nature Documentation, IBM Quantum, qiskit-community.github.io/qiskit-nature/apidocs/qiskit_nature.second_q.mappers.html.

"StatevectorEstimator." IBM Quantum Documentation, quantum.cloud.ibm.com/docs/en/api/qiskit/qiskit.primitives.StatevectorEstimator.

"VQE." IBM Quantum Documentation, v0.46, quantum.cloud.ibm.com/docs/en/api/qiskit/0.46/qiskit.algorithms.VQE.

Challa, Keerthan. "VQE-EthyleneOxide-6." GitHub, 2026, github.com/KeerCode/VQE-ethyleneOxide-6.

Breast Cancer Awareness: What Are the Things Young Girls and Women Can Do Every Day to Prevent Breast Cancer? By Adora Liao

Abstract

Breast cancer is the second leading cause of women's death. According to WHO's breast cancer statistics, there were 2.4 million breast cancer cases in 2024, and 694,000 deaths globally. Early diagnosis of breast cancer is a crucial way to prevent cancer cells from spreading to other organ systems. When individuals do not recognize the mild symptoms, they often experience a prolonged "appraisal interval", the span from noticing the symptoms to deciding to go see a doctor. The lack of awareness may lead to lower chances of full recovery. The objective of this research is to explore the relationship between dietary pattern and breast cancer (BC) risk rate, with the goal of raising awareness of breast cancer and encouraging young girls and women to adopt a healthy lifestyle in order to reduce the risk of breast cancer. This study collected the data from different research papers and summarized them into two categories: genetic predisposition, and personal lifestyle choices. Furthermore, the study identified three accessible methods that girls and young women can easily achieve in daily life to significantly lower the incident rate of getting breast cancer and increase their well-being. The results show the significant influence of diet and genetics on breast cancer. Genetic mutation is strongly connected to the risk rate of breast cancer. However, implementing a healthy diet pattern and consuming sufficient vitamin D are also crucial to the prevention of breast cancer. and the importance of making health-conscious lifestyle choices. Breast cancer is a prevailing disease around the world and can affect many women's lives. Early breast cancer detection with proper treatments could profoundly reduce the death rate in the long-term (Wang 1). Since breast cancer is a common cancer in women, early diagnosis and healthy diet patterns play important roles in the recovery processes and the prevention of breast cancer.

Introduction

Breast cancer (BC) is a disease in which cancer cells grow rapidly and lead to the development of tumours. Globally, over 1.5 million women are diagnosed with breast cancer every year (Sun 1). The incidence of BC has shown an upward trend from 2012 to 2021, with a 1% annual increase (Giaquinto 1). According to Giaquinto et al. (2024), in 2024, around 310,720 women in the US were diagnosed with invasive breast cancer, and 42,250 women will die from BC. Approximately 13% of the women (one in eight) in the US will be diagnosed with invasive breast cancer, and 2% (one in 43) will die from BC.

Moreover, young women (below 40 years old) have a higher risk of having tumours with negative clinicopathologic features, such as higher histological grade, more lymph node positivity, lower estrogen receptor (ER) positivity, and higher rates of Her2/neu overexpression (Radecka 1). In addition, although the incidence of BC is much less common under the age of 40 than the counterpart of older women, young women's diagnosis of BC tends to present at later stages, and be of a more aggressive subtype. For example, according to Zhu et al.(2023),

triple-negative or HER2-positive breast cancer. The BC patients under 40 years old also have a higher risk of having a poorer prognosis, such as poor tumour differentiation, increased Ki-67 expression, and more extensive lymph node involvement, compared to women >50 years of age. These collective studies reveal that young breast cancer patients can be correlated with an increased susceptibility to experience more aggressive symptoms than older patients.

Regular mammography and early detection significantly improve the chances of successful treatment and survival. According to Ginsburg, O. et al (2020), a large proportion of breast cancer patients are diagnosed with later-stage disease. Therefore, advocating regular medical examinations and early diagnosis will improve outcomes for all BC patients.

The objective of this research is to explore the factors that lead to an increase in breast cancer risk and the ways people can be assessed easily to reduce the possibility of getting breast cancer, intending to promote young women and girls to raise breast cancer awareness and choose a healthier lifestyle pattern. This paper has 4 main sections. Section 1 introduces breast cancer and highlights the significant impact of breast cancer around the world. Section 2 reveals two primary factors that strongly contribute to the incidence of breast cancer. Section 3 lists three measures in everyday life that can help alleviate or reduce the risk of breast cancer. Section 4 concludes the paper and further substantiates the need to enhance breast cancer awareness and promote a healthy lifestyle among the public.

Dominant Factors affecting Breast Cancer Risk

Genetic Predisposition

Certain genetic mutations are strongly connected to the risk of breast cancer. In particular, mutations in the BRCA1 and BRCA2 genes play a major role in tumour development. BRCA1 and BRCA2 are anti-oncogenes associated with breast cancer risk and are located on chromosomes 17q21 and 13q12, respectively. According to a recent study, female carriers of mutations in BRCA1 or BRCA2 have a lifetime risk of breast cancer of 50%–85%. (Shiovitz 1292). In general, BRCA1 and BRCA2 are the strongest susceptibility genes for BC. Mutations in these two specific genes account for up to 90% of hereditary BC cases. BRCA1/2 mutations are inherited in an autosomal dominant manner. BRCA1 mutations can disrupt cell cycle checkpoints, lead to abnormal centrosome duplication, genetic instability and apoptosis. BRCA2-associated breast cancers are more likely to be high-grade invasive ductal carcinomas with a luminal phenotype. According to a meta-analysis by Chen, the breast cancer risk ratio in women older than 70 years carrying BRCA1 and BRCA2 mutations was 57% and 49%, respectively. Other associated genes, such as HER2 (c-erbB1), c-Myc, and Ras, are also linked to increased breast cancer risk. Women with mutations in BRCA1/2 should be aware of BC and conduct breast self-exams regularly starting at age 18, as recommended by the European Society for Medical Oncology (ESMO) and the National Comprehensive Cancer Network (NCCN) guidelines.

Lifestyle Choices

Lifestyle factors play a crucial role in the development of breast cancer. According to Max Dieterich's study on the impact of lifestyle factors on breast cancer risk, several behaviours and conditions—including obesity, alcohol consumption, shift work factors, use of antiperspirants, physical inactivity, hormone therapy after menopause, diabetes mellitus, and stress—are associated with the increased risk of breast cancer. In daily life, our diet, body training, and awareness of breast cancer are vital to the prevention of breast cancer. Integrating nutrition into both preventive and post-diagnosis care is what people can apply to reduce the risk of getting breast cancer. Several unhealthy lifestyle patterns can increase the risk of BC.

The consumption of alcohol is a well-known factor that can result in negative impacts on the risk rate of BC. "A 10% increased risk of BC has been reported for every additional 10g ethanol consumed daily."(Dieterich 408). The use of antiperspirants has contributed to BC risk, as the ingredients in antiperspirants, such as parabens, can lead to rising BC case numbers in the upper outer quadrant.

Being overweight or obese is also a well-established risk factor for breast cancer due to its effects on inflammation, hormone levels, factors that regulate cell growth, and others. Combined continuous hormone therapy after menopause for women has a significant impact on BC risk rate. Other lifestyle factors like diabetes mellitus, exposure to chemicals, shift work factors, and breast implants also influence the risk rate of BC.

Risk reduction strategies to prevent breast cancer

Healthy diet

Applying a healthy diet is a crucial factor that reduces the risk of breast cancer. High consumption of certain foods is associated with a significantly decreased risk. A plant-based diet may reduce breast cancer risk by regulating estrogen levels and reducing inflammation. (How Diet Impacts Breast Cancer Risk and Outcomes | Breast Cancer Research Foundation) According to Do, T.M. et al. (2024), dark green vegetables, legumes, and starchy vegetables are foods that can lower the risk of breast cancer. Polyphenols found in fruits and vegetables may antagonize estrogen pathways, inhibiting cancer cell proliferation. (Pontillo 4). However, the impact of fibre has been controversial. Some research suggests that consuming fibre can help control blood sugar, lower estrogen levels, and influence the composition of the gut microbiome (affecting risk by altering estrogen metabolism). On the other hand, some reported no significant associations between dietary fibre intake and BC risk. Other factors, like high intake of dietary fats and soy isoflavones, may be linked to the development of breast cancer.

Vitamin D

Vitamin D deficiency is considered a negative prognostic factor in women with breast cancer. According to Atoum M et al., vitamin D receptor (VDR) genes contribute to the increased risk of BC. Vitamin D receptor genes regulated by vitamin D play important roles in

the mammary gland by regulating calcium transport during lactation, hormone differentiation, and milk production. Vitamin D deficiency is also common in BC patients. In a recent study, 95.6% of the BC patients had vitamin D deficiency. (Shaukat 646) Moreover, BC patients with vitamin D deficiency are identified to have a more aggressive molecular phenotype and worse prognostic indicators. Practice guidelines recommend vitamin D and calcium supplementation in BC patients to improve bone density and minimize fracture risk. Therefore, maintaining adequate vitamin D intake is important, as deficiency may negatively influence BC progression.

Mediterranean diet

Lastly, adherence to a Mediterranean diet is associated with a decreased risk of breast cancer. The Mediterranean diet is a dietary pattern of consuming food like fruits, vegetables, whole grains, and lean proteins. Consuming fruits and vegetables, such as cruciferous vegetables and citrus fruits, may reduce the risk of getting BC. Recent studies have shown that the Mediterranean diet is linked with gene expression variation and the regulation of molecular pathways involved in cell proliferation, apoptosis, and angiogenesis. For instance, polyphenols present in fruits and vegetables can regulate the activity of genes associated with cell survival and metastasis. In addition, tyrosol and hydroxytyrosol present in extra-virgin olive oil have the ability to reduce glutathione levels, which is beneficial for breast cancer prevention. According to Claudia Reytor-González et al. (2025), bioactive components in the Mediterranean diet, such as the monounsaturated fatty acids in olive oil and antioxidants found in fruits and vegetables, are significantly associated with BC treatment and prevention. The monounsaturated fatty acids and phytochemicals help reduce inflammation and alleviate stress in the human body, processes that profoundly influence the progression and development of BC. There is an abundance of natural, nutrient-rich products in the Mediterranean region that studies show help prevent cancer from developing and spreading, such as pomegranate, edible mushrooms, marine algae, curcumin, whole-grain cereals, citrus fruit, grapes, and cucumber. (Tsai 4) According to Healthcare (Basel). (2025), implementing a Mediterranean diet can lower the BC risk by 12% per unit increase. These collective studies indicate that the Mediterranean diet can effectively help reduce the risk of getting BC.

Conclusion

According to the Breast Cancer Research Foundation, up to 50 to 70 per cent of breast cancer cases can be prevented (“How Diet impacts Breast Cancer Risk and Outcomes”, 2024), depending on when people make healthy lifestyle changes. Healthy lifestyles, including proper diet and body training, are associated with lower levels of inflammation, and certain healthy diet patterns also contribute to the reduction of breast cancer risk. Although inherited genes have a strong influence on the development of breast cancer and tumours, following a plant-based diet, maintaining a healthy body weight, limiting alcohol consumption, and choosing a healthy lifestyle allows women to reduce the possibility of getting breast cancer. As for daily methods for young women to adopt, consuming adequate vitamin D, following the Mediterranean diet, and

adopting a plant-based diet are not only accessible and easy for the public to implement but also significantly reduce the risk of BC.

Works Cited

"Breast Cancer Diet & Nutrition." Breast Cancer Research Foundation, https://www.bcrf.org/about-breast-cancer/breast-cancer-diet-nutrition/.

Sun, Yi-Sheng, et al. "Risk factors and preventions of breast cancer." International journal of biological sciences 13.11 (2017): 1387.

Dieterich, Max, et al. "Influence of lifestyle factors on breast cancer risk." Breast care 9.6 (2014): 407-414.

Pontillo, Martina, et al. "Dietary interventions for breast cancer prevention: Exploring the role of nutrition in primary and tertiary prevention strategies." Healthcare. Vol. 13. No. 4. MDPI, 2025.

Do, Tam M., et al. "Association between dietary factors and breast cancer risk: a matched case-control study in Vietnam." BMC cancer 24.1 (2024): 1224.

Mourouti, Niki, et al. "Diet and breast cancer: a systematic review." International journal of food sciences and nutrition 66.1 (2015): 1-42.

"Breast Cancer." World Health Organization, 3 July. 2026, https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Wang, Lulu. "Early diagnosis of breast cancer." Sensors 17.7 (2017): 1572.

Shiovitz, Stacey, and Larissa A. Korde. "Genetics of breast cancer: a topic in evolution." Annals of Oncology 26.7 (2015): 1291-1299.

Giaquinto, Angela N., et al. "Breast cancer statistics 2024." CA: a cancer journal for clinicians 74.6 (2024): 477-495.

Radecka, Barbara, and Maria Litwiniuk. "Breast cancer in young women." Ginekologia polska 87.9 (2016): 659-663.

Zhu, Jie Wei, et al. "What is known about breast cancer in young women?." Cancers 15.6 (2023): 1917.

Ginsburg, Ophira, et al. "Breast cancer early detection: A phased approach to implementation." Cancer 126 (2020): 2379-2393.

Mehrgou, Amir, and Mansoureh Akouchekian. "The importance of BRCA1 and BRCA2 genes mutations in breast cancer development." Medical journal of the Islamic Republic of Iran 30 (2016): 369.

Atoum, Manar, and Foad Alzoughool. "Vitamin D and breast cancer: latest evidence and future steps." Breast cancer: basic and clinical research 11 (2017): 1178223417749816.

Williams, Jasmaine D., et al. "Tumor autonomous effects of vitamin D deficiency promote breast cancer metastasis." Endocrinology 157.4 (2016): 1341-1347.

Shaukat, Noureen, et al. "Association between vitamin D deficiency and breast cancer." Pakistan journal of medical sciences 33.3 (2017): 645.

Reytor-González, Claudia, et al. "Mediterranean diet and breast cancer: A narrative review." Medwave 25.2 (2025): e3027.

Tsai, Hsueh-Han, et al. "The risk of breast cancer between western and mediterranean dietary patterns." Nutrients 15.9 (2023): 2057.

The Aftermath of The 2008 Economic Crisis Regarding Firms, Households, and Financial Institutions By Dan Vu

Introduction

The 2008 Economic Crisis, also known as the Great Recession, was one of the most significant economic downturns in American history since the Great Depression. It was triggered by the bursting of the US housing bubble, driven by subprime mortgages (mortgages given to people with poor credit scores), toxic assets (financial assets whose market value had collapsed because they were backed by risky mortgages), surges in defaults (when borrowers fail to make required loan repayments), and failures in financial regulation. When major banks such as Lehman Brothers fell, the global credit market froze as banks became unwilling to lend due to fears of losses, resulting in widespread unemployment and financial hardship worldwide. Ultimately, the government had to intervene by purchasing a substantial number of toxic assets, thereby artificially boosting the economy until it returned to normal. While this intervention and bailout eventually stabilized the economy, the consequences left a lasting impact worldwide (Duignan). All data used in this essay were primarily drawn from the Federal Reserve Bank of St Louis. This essay aims to explore the aftermath of the 2008 economic crisis in relation to firms, households, and financial institutions using key economic indicators. By comparing these three groups, it demonstrates that although government intervention stabilized the financial system, the recovery was uneven, with households experiencing the slowest and most persistent consequences. Understanding these differences provides a clearer evaluation of both the effectiveness and the limitations of the government's response.

Literature review

There have been many debates regarding how the economic crisis was dealt with, whether or not it was properly handled, and whether its impact was minimized. Some places attribute the cause of the financial crisis to reckless lending and financial mismanagement by the banks, while others blame weak regulation and government policy failures. One such example is Bernie Sanders, a U.S. Senator, who blames banks for the crisis, stating that “One of the biggest mistakes our government made after the financial crisis was not prosecuting the people responsible for the greed, recklessness and illegal behaviour that crashed our economy and ruined the lives of millions of Americans”. On the other hand Nigel Farage, British Member of Parliament blamed the lack of regulation “The banking collapse was caused, more than anything, by bad government policy and the total failure of bad regulation, rather than by greed.” (Ozili).

Many people have differing opinions on the actions and interventions the government has taken to minimize the damage and bring the economy back from its downturn. While some compliment the government's regulation, many have expressed unease at the fact that government intervention might cause more harm than good, as the economy will gradually be dependent on them. An example of these differing viewpoints is Stephen A. Schwarzman, an American businessman who support the regulation, “In a financial crisis, only the Fed, as the

lender of last resort, might stand between our economy and financial catastrophe." and Ludwig Von Mises, who said, "Every government intervention [in the marketplace] creates unintended consequences, which lead to calls for further government interventions." (Ozili). All these contrasting arguments show that economists are still unable to reach a clear conclusion on whether or not the 2008 economic crash was truly solved, and that this issue, in some ways, is still prevalent today.

Impact of the 2008 economic crisis

i) Impact of the 2008 Crisis on Banks.

One of the major players and at the same time one of the most affected groups during the economic crisis was the commercial banks. During the crisis, two of the main aspects of the bank that were affected were loans and leases - the amount of money banks were lending to businesses - and the banking reserves - the minimum amount of money the bank must have on hand at any given time.

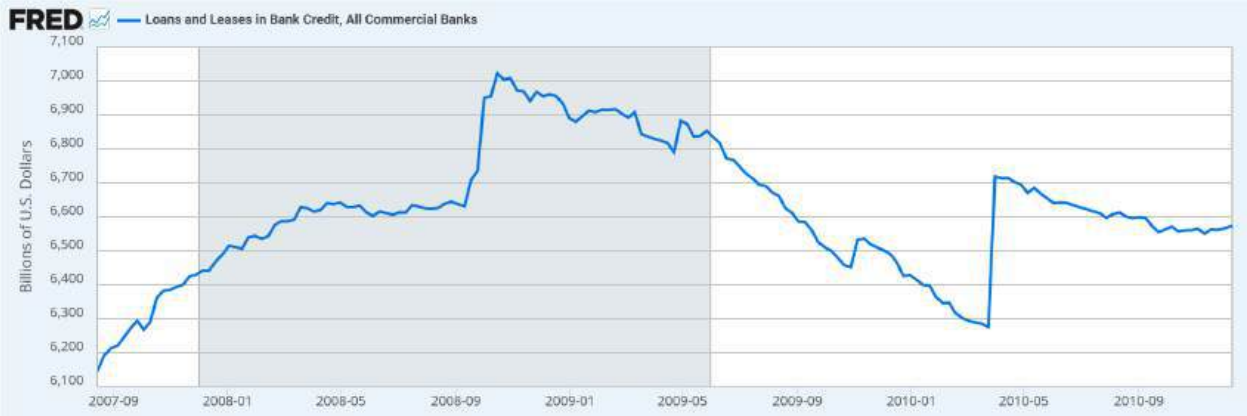


Fig 1. Loans and Leases in Bank Credit, All Commercial Banks, from September 2007 to September 2010 (Board of Governors of the Federal Reserve System)

Figure 1 shows the total value of loans and leases given by commercial banks to firms and households. Looking at the data before the start of the 2008 Crisis, the amount of loans Banks were steadily increasing before spiking on October 15; afterwards, they rapidly decreased until the government injected more money into banks, leading to the loans and leases returning to normal for a certain amount of time in 2010. While the reason loans and leases spiked on October 15 and not September 15 is a bit unclear, it could be explained by the fact that banks took time to start and initiate a credit crunch (period during which banks sharply reduce lending because they perceive lending as too risky). Even so, the surge in loans was temporary, and by 2009, loan growth had reduced dramatically. This decline was likely due to banks fearing rising defaults, resulting in them pulling back from risky lending. Instead, the bank started focusing on hoarding reserves at the Federal Reserve, which offers a much smaller but safer and more secure

option for profit. This contraction ended up deepening unemployment, slowing down business investment, and prolonging the recession, showing how the collapse of bank lending during 2008 was not only a symptom of the crisis but also one of the driving causes that made it so bad.

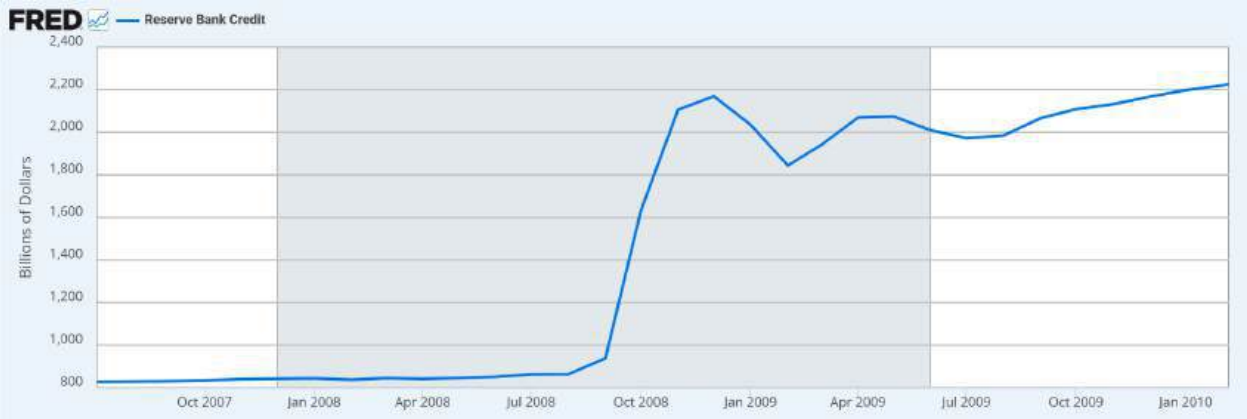


Fig 2. Reserve Bank Credit from October 2007 to January 2010 (Board of Governors of the Federal Reserve System)

A broader perspective on the 2008 crisis can be seen in Figure 2, where banks rapidly increased their Reserve Bank credit in the span of a year. Before the crisis, banking reserves usually straddled a bit above eight hundred billion dollars, which means that if a bank run were to happen, banks would not have enough money on hand to give out to all the people. This was one of the reasons why the 2008 crisis was so detrimental, as banks simply were not ready to handle the consequences of a bank panic when the defaults started rising. However, when TARP was enacted, banks had more liquid assets as the Federal Reserve massively expanded banks' reserves through emergency lending (discount window and special liquidity facilities). However, since banks were still reluctant to make loans, they maintained the credit crunch all the way till 2010 before loans were given out again. In the end, while the rise in reserves prevented systemic collapse in the financial system by providing liquidity (easily accessible money) to banks, it also reinforced the post-credit crunch as banks continued to prioritise fixing their own assets (Singh).

ii) Impact of the 2008 Crisis on Firms and Businesses.

The next group that was affected by the economic crisis is the firms across different industries. For businesses, the primary reason for their impact was reduced access to credit and declining consumer demand. The aspects that were affected include GDP, which measures the overall output of goods and services, and the delinquency rate on business loans, which captures how many firms failed to meet their debt obligations.

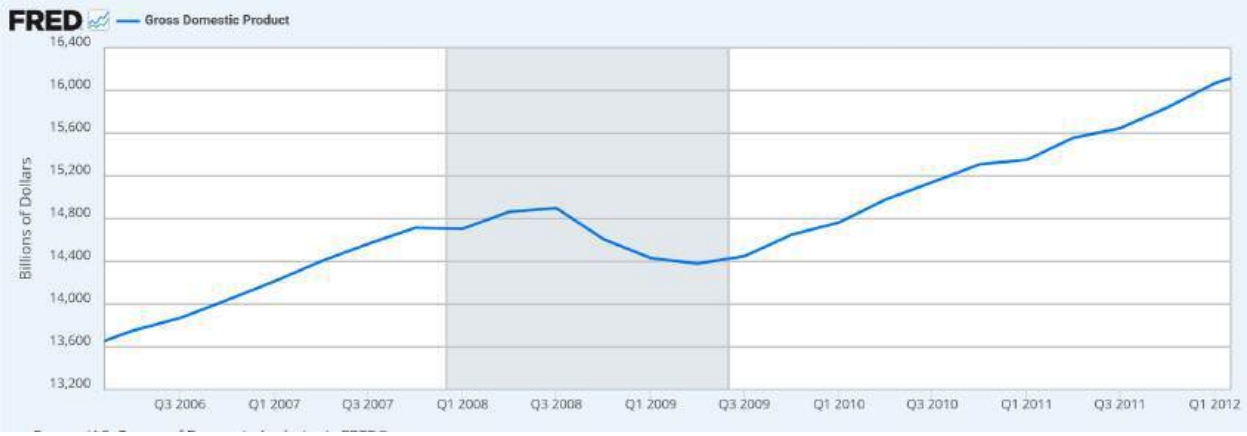


Fig 3. Gross Domestic Product (GDP) from the Third quarter of 2006 to the first quarter of 2012 (U.S. Bureau of Economic Analysis)

This data shows the gross domestic product over time. Looking at the data, we can see there was a small dip in GDP during the Third quarter of 2008, which represented falling production and demand across the economy at that time. As households reduced their consumption and banks reduced access to credit, firms and businesses saw revenue and sales collapse. Despite the dip in data looking small, a fall of 4 to 5 percent in GDP means that the economy was producing 4 to 5 percent less output every year. This means that hundreds of billions of USD were being lost, leading to massive layoffs that further decreased consumption, creating a vicious cycle. Furthermore, because GDP usually grows steadily, even a small dip can represent a historic collapse in output and confidence, negatively affecting the economy in the long run. The fall in GDP also shows how deeply the firms were affected, as hundreds were likely forced to close or contract. This contraction not only lowered employment but also slowed the economic recovery, as businesses became significantly more cautious.

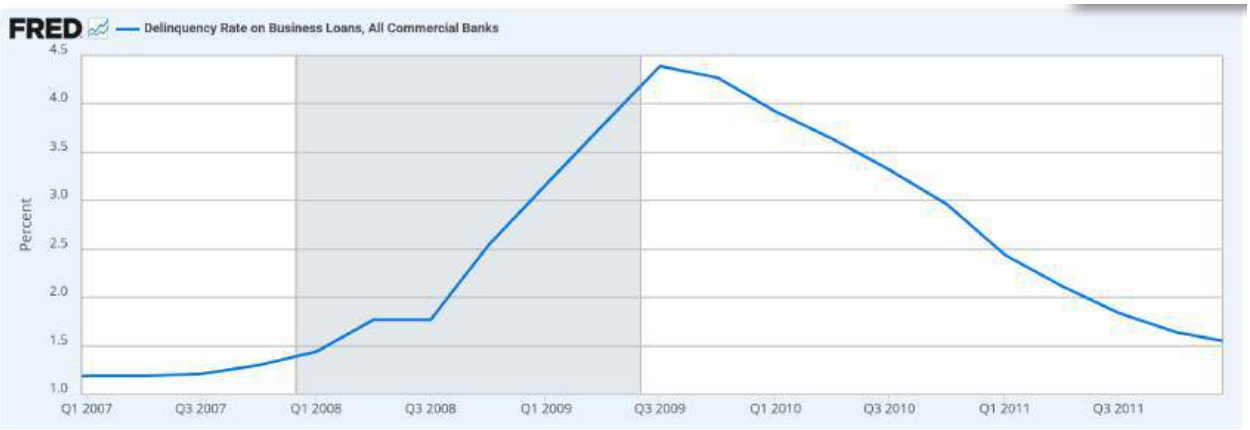


Fig 4. Delinquency Rate on Business Loans, All Commercial Banks from Quarter 1 of 2007 to Quarter 3 of 2011 (Board of Governors of the Federal Reserve System).

Another key measure of the crisis's impact on firms was the delinquency rate of business loans, which records how often companies failed to repay debt on time (Board of Governors of the Federal Reserve System (US), "Delinquency Rate on Business Loans, All Commercial Banks"). During the 2008 crisis, the delinquency rate surged and quadrupled as revenue collapsed and credit conditions tightened. With demand falling significantly along with the credit crunch, many small and mid-sized businesses could not meet their loan obligations, leading to defaults and bankruptcies. This spike in the delinquency rate shows how the credit contraction prolonged the recession. Banks deemed businesses a risky option for lending credit, further deepening the cycle of contraction. For many firms, this meant they had to close, reduce investment, and downsize, slowing the recovery far after the official end of the recession.

iii) Impact of the 2008 Crisis on Households.

Households were also deeply affected by the crisis, experiencing some of the most visible and long-lasting consequences of all parties. The labor market was severely disrupted, with unemployment rates rising sharply and labor force participation declining as many people gave up searching for jobs and simply ended up homeless. Housing, one of the most important areas for households, was especially hard hit, as homeownership rates fell sharply. This shows how the financial crisis not only disrupted day-to-day living for millions of families but also affected the employment, spending, and housing stability of all families near or inside the US in the long term.

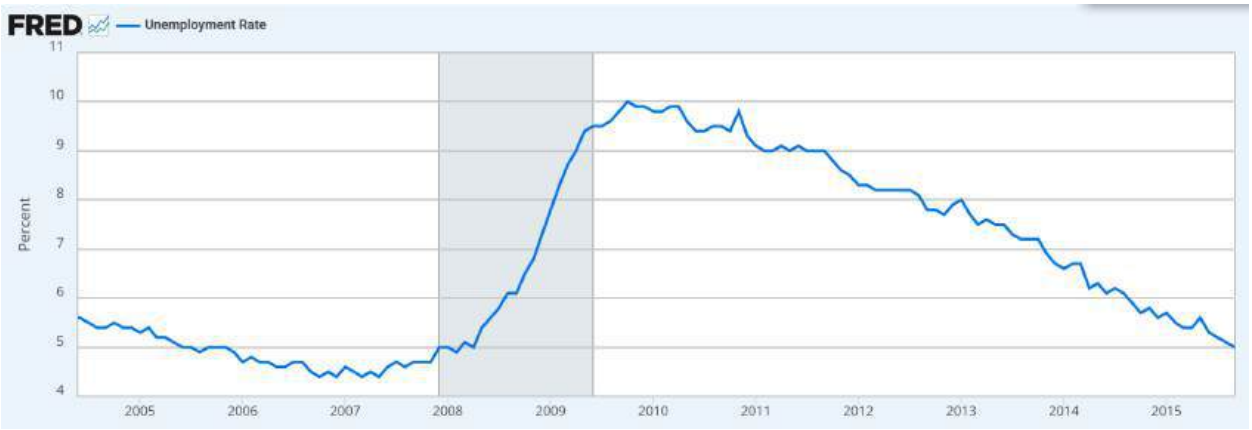


Fig 5. Unemployment Rate from 2005 to 2015 (U.S. Bureau of Labor Statistics)

The labor market was devastated during the 2008 crisis, with the unemployment rate doubling from 5 to 10% in the span of 2 years, one of the highest levels of unemployment in decades. This means that millions of workers are losing their jobs in a short amount of time, most particularly in the industries tied to housing, construction, manufacturing, and finance. For most families, unemployment meant the sudden loss of income, health insurance, and stability, forcing them to rely on what little savings they had and high-interest debt. The long duration of unemployment was also damaging, as many workers remained jobless for a long time, leading to

skill erosion, psychological strain, and a permanent scar on the labor market, as many people dropped out of the labor force. Even as GDP began to recover, as seen in Fig. 3, the unemployment rate remained elevated for several years, showing that the human cost outlasted the financial cost during the crisis.

Fig 6. Labor Force Participation Rate from July 2006 to January 2014 (U.S. Bureau of Labor Statistics)

Alongside the unemployment rate, there is also the labor force participation rate, which represents people who are employed and unemployed. During the crisis, the labor force participation rate dropped sharply for many years straight, seemingly unable to recover. Many discouraged workers stopped looking for jobs altogether after repeated failures to find employment. This ends up masking the true effect of the crisis, as the official unemployment rate only counts when a person is actively seeking work. The significant drop in the labour force participation rate, creates a deceptive story that many people were reemployed, despite the data merely showing that many workers exited from the labour force. The lower participation rate can also represent the fact that many individuals went back to school or retired early, or perhaps had more informal employment. The declining labor force means that even as the economy slowly expanded, fewer people shared in the recovery. This reduction in workforce engagement reduced household income, weakened consumer spending, and created a long-lasting effect on inequality in class and economic potential.

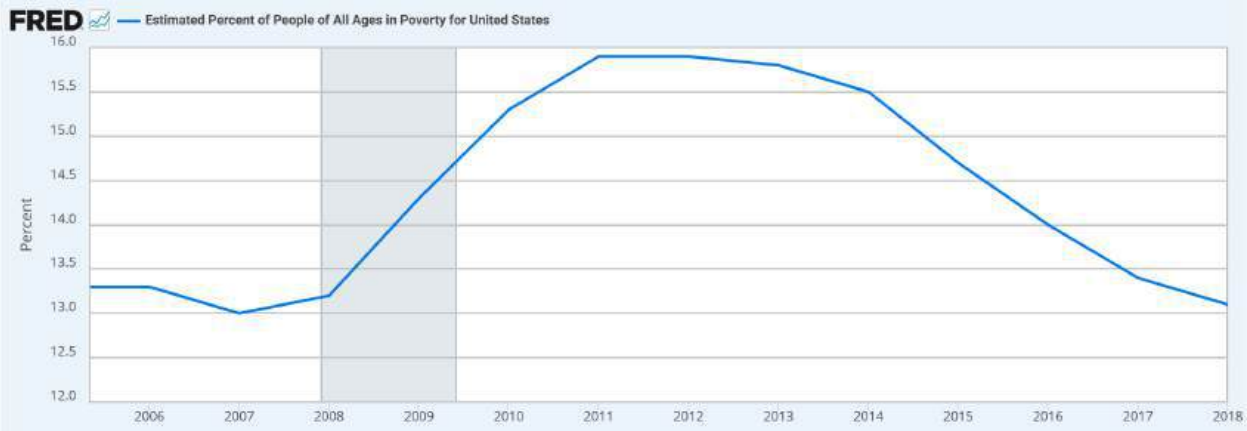


Fig 7. Estimated Percent of People of All Ages in Poverty for the United States (U.S. Census Bureau)

The poverty rate provides another lens into the household impact of the crisis. According to Fig. 7, the share of Americans living in poverty was 13 percent before reaching 16 percent in 2011, once again being one of the highest rates for multiple decades. Because the poverty line was around 21,000 dollars for a family of four in 2008, the increase in the poverty rate was substantial. This spike reflected the combination of mass unemployment, reduced earnings, and credit crunches that wiped out household wealth. Unlike financial institutions that received government support, many families remained under pressure for years, only finally starting to recover in 2014. The slow decline in poverty illustrates how households experienced a much more drawn-out recovery compared to banks and firms, highlighting how the Great Recession had a larger burden on poorer households.

Government policy

In response to the 2008 economic crisis, the government put in place multiple policies to restore confidence among consumers and banks. Authorized on October 20th, 2008, the most significant government policy was TARP (Troubled Asset Relief Program), which allowed for an injection of capital directly into banks by purchasing preferred stocks: the purchaser of preferred stocks receives a larger percentage of money from banks but is not allowed to directly influence the governance and decision-making of those banks (Ganti). This helped prevent widespread bank failures as banks now had more money on hand. In total, 443 billion dollars were spent during the program. Although controversial for appearing to favor Wall Street over struggling households, TARP succeeded in stabilizing the financial system, and most of its funds were eventually repaid. An article by Troy Segal, however, has pointed out that TARP acted as a reward for bad behavior, sending a message of "act irresponsibly, and we'll help you out" (Segal).

Along with this, the Federal Reserve also applied Quantitative Easing (QE), purchasing large quantities of government bonds and mortgage-backed securities to lower long-term interest rates, pushing banks to give out more loans to make more money. Lower rates made it cheaper for households to borrow for mortgages and for businesses to take loans for investment. Together

with numerous other policies, America managed to avoid systemic collapse. However, the uneven speed of recovery, in which firms recovered significantly faster than households, highlighted the limits of the government response.

Conclusion

In conclusion, the 2008 economic crisis left a lasting mark on every part of the economy. For banks, it meant a sudden collapse in lending, rising delinquencies, and a shift toward holding reserves and safe assets rather than financing growth. For firms, the sharp decline in GDP and soaring delinquency rates on business loans forced widespread layoffs, closures, and a long pause in investment. Households suffered the heaviest burden as unemployment doubled, labor force participation fell significantly, poverty rates climbed, and millions lost their homes. Despite the government policy restoring the economy, the growth was uneven and slow. In the end, the 2008 crisis shows both the necessity and the limitations of policy intervention as the Great Recession forever left its scar on the American economy.

Works Cited

Board of Governors of the Federal Reserve System (US). "Bank Credit, All Commercial Banks." FRED, Federal Reserve Bank of St. Louis, 3 Jan. 1973, fred.stlouisfed.org/series/TOTBKCR. Accessed 29 Sept. 2025.

---. "Delinquency Rate on Business Loans, All Commercial Banks." FRED, Federal Reserve Bank of St. Louis, 1 Jan. 1987, fred.stlouisfed.org/series/DRBLACBS. Accessed 29 Sept. 2025.

---. "Loans and Leases in Bank Credit, All Commercial Banks." FRED, Federal Reserve Bank of St. Louis, 3 Jan. 1973, fred.stlouisfed.org/series/TOTLL. Accessed 29 Sept. 2025.

---. "Real Estate Loans: Commercial Real Estate Loans, All Commercial Banks." FRED, Federal Reserve Bank of St. Louis, 2 June 2004, fred.stlouisfed.org/series/CREACBW027SBOG. Accessed 29 Sept. 2025.

---. "Reserve Bank Credit." Stlouisfed.org, 2025, fred.stlouisfed.org/series/RSBKCRNS. Accessed 29 Sept. 2025.

Duignan, Brian. "Financial Crisis of 2007–08." Britannica, 3 July 2024, www.britannica.com/money/financial-crisis-of-2007-2008. Accessed 29 Sept. 2025.

FRED. "Unemployment Rate." Stlouisfed.org, 6 June 2025, fred.stlouisfed.org/series/UNRATE. Accessed 29 Sept. 2025.

Ganti, Akhilesh. "Preferred Stock Definition." Investopedia, 4 Sept. 2021, www.investopedia.com/terms/p/preferredstock.asp. Accessed 29 Sept. 2025.

Ozili, Peterson. Munich Personal RePEc Archive 100 Quotes from the Global Financial Crisis: Lessons for the Future. 30 Dec. 2022, mpra.ub.uni-muenchen.de/115824/1/MPRA_paper_115824.pdf. Accessed 22 Sept. 2025.

Segal, Troy. "Troubled Asset Relief Program (TARP), What It Was, How It Worked." Investopedia, 29 Sept. 2020, www.investopedia.com/terms/t/troubled-asset-relief-program-tarp.asp. Accessed 29 Sept. 2025.

Singh, Manoj. "The 2008 Financial Crisis Explained." Investopedia, Investopedia, 25 Aug. 2024, www.investopedia.com/articles/economics/09/financial-crisis-review.asp. Accessed 29 Sept. 2025.

U.S. Bureau of Economic Analysis. "Gross Domestic Product." Stlouisfed.org, 30 Apr. 2025, fred.stlouisfed.org/series/GDP. Accessed 29 Sept. 2025.

U.S. Bureau of Labor Statistics. "Labor Force Participation Rate." FRED, Federal Reserve Bank of St. Louis, 2024, fred.stlouisfed.org/series/CIVPART. Accessed 29 Sept. 2025.

U.S. Census Bureau. "Estimated Percent of People of All Ages in Poverty for the United States." Stlouisfed.org, 2018, fred.stlouisfed.org/series/PPAAUS00000A156NCEN. Accessed 29 Sept. 2025.

A Lost Page in History: Identifying the First Asian American US Senate Page
By Jayden Kim, Advisor By Katherine A. Scott, Ph.D.

Abstract

The United States Senate Page Program, established in 1829, is one of the oldest and most significant Congressional employment programs for youth, allowing high school juniors ages 16 and 17 to serve the United States Senate for one academic semester. Despite the US Senate's recognition of major milestones such as the appointment of the first African American and female Senate Pages, there is no institutional recognition for the first Asian American Senate Page or the history of Asian Americans in the Senate Page Program. This study identified the first Asian American Senate Page through the triangulation of archival sources. First, the *Report of the Secretary of the Senate* was used to filter potential Asian American Senate Pages through a surname-based identification (SBI) strategy. Second, using Senate Page yearbooks, participation in the program was confirmed and identifying details were refined (e.g., photographic and contextual details). Third, for each candidate, historical newspaper databases were searched to confirm the Asian American identity as well as other contextual details such as whether the individual was the first Asian American Senate Page. Through this process, Spencer Chen, a Chinese American student from Lincoln High School in Portland, Oregon appointed by Senator Maurine Neuberger in 1965, was identified as the first Asian American Senate Page. A historic Senate Page from the times of major Civil Rights legislation, Chen symbolizes a previously unrecognized "first step" towards racial integration for Asian Americans in the Senate Page Program. This study illustrates how recovering overlooked histories can help shape a fuller and more inclusive institutional narrative.

Introduction

The United States Senate Page Program has existed for almost 200 years and continues to be one of the most long-standing and substantive youth service programs in the United States Congress. In 1829, Senator Daniel Webster appointed Grafton Hanson to become the first formally designated Senate Page (United States Senate Page Program). Currently, each semester, 30 high school juniors are selected to work and learn in Washington, D.C. through a six-month employment at the United States Senate. Each student is sponsored and appointed by a United States senator. During the time they are employed as a Senate Page, students carry out many day-to-day tasks of the Senate such as delivering communications, carrying papers/documents, developing documentation, and supporting senators while the Senate is in session (see fig. 1). Every four years, Senate Pages participate in the presidential election certification process by carrying the large mahogany boxes with the electoral vote certificates into the joint session of Congress, which has been an important tradition in maintaining the peaceful transfer of power. Notable former Senate Pages include former US Vice President Spiro T. Agnew, US Supreme Court Associate Justice Neil Gorsuch, Governor Abigail Spanberger, Senator Michael Bennet, and Senator Mike Lee (see fig. 2).

Fig. 1. Senate Pages working on the Senate floor rostrums during a roll-call vote. Screenshot from "US Senate," C-SPAN2.

Fig. 2. Notable former US Senate Pages: **(a)** Spiro Agnew, Vice President in the Nixon Administration; **(b)** Neil Gorsuch, Associate Justice; **(c)** Abigail Spanberger, first female governor of Virginia; **(d)** Senator Michael F. Bennet; and **(e)** Senator Mike Lee. Sources, respectively: *Encyclopaedia Britannica*; photograph by Franz Jantzen, Collection of the Supreme Court of the United States, via Wikimedia Commons; Appalachian Regional Commission; United States Senate Historical Office; and photograph by Kevin Dietsch/Getty Images, via *Wisconsin Examiner*.

Because Senate Pages represent employment selections made by US senators, federal lawmakers in the United States, the Senate Page Program can demonstrate how representation and inclusion in the US Senate has developed over time. A historical report by the US Senate notes that the US Senate Page Program, which has for many years been made up primarily of males with a near-exclusive representation by white individuals, has begun to include greater diversity since the middle of the twentieth century (United States Senate, "Historical Overview"). The appointment of Pages by members of Congress was viewed not only as an opportunity for students to be employed, but also as an issue of whether the institution of the Senate was fair and accessible to all people (United States Senate Historical Office).

Throughout the program's history, many "firsts" have been recognized, such as the

appointment of Andrew Slade as the first African American Senate Page in 1869, Emma Madeen as the first female Senate Page in 1907, and John Lopez as the first Puerto Rican Senate Page in 1966 (Amer). The presence of these historically marginalized groups in political decision-making bodies can signal broader social and political change (Clayton et al.). However, whether and how those "firsts" are documented, remembered, and publicly recognized depends on institutional recordkeeping and the construction of official narratives. If records were not kept and official narratives were not constructed, there lie possibilities of hidden or forgotten "firsts." Official Senate historic records have recognized a number of significant events in Page representation (e.g., the first African-American and the first female Senate Pages). Yet it appears that the same level of formal acknowledgment has not been provided to all racial/ethnic groups.

The *Report of the Secretary of the Senate*, the semiannual Senate employment and expenditure record, lists Senate employees, including Senate Pages, by name and position but does not include racial or ethnic identifiers (Petersen; Jones and Triguero Roura). As a result, to develop the methodology for finding hidden or forgotten "firsts," it is required to reconstruct history through indirect sources. This challenge of being hidden and forgotten in historical records is especially significant for Asian Americans. Political scientists and historians have argued that Asian Americans are often underrepresented in formal political institutions and, at the same time, rendered "invisible" in broader narratives of US politics (Kim 118; Wong et al. 45–70). The invisibility is reinforced when institutions fail to record or publicize milestones in Asian American participation. This often leaves the impression that Asian Americans were largely absent from elite political spaces.

Representation and sense of belonging are related. Research shows that visibility and representation is critical to create feelings of recognition, value, and legitimacy because people are more likely to feel included if individuals they associate with are seen in institutional spaces (Gay 717). In addition, when individuals are recognized, such as through officially acknowledging a "first," the recognition can influence how individuals understand their connection to institutions.

Asian Americans are the fastest-growing racial group in the US (Budiman and Ruiz). Over the past century, not only has the Asian American population grown, but there have also been nine senators of Asian American descent, including the first Asian American senator, Hiram Fong; the first female Asian American senator, Mazie Hirono; and the first Korean American senator, Andy Kim.

Despite these understandings and advancements of Asian Americans, there is no record recognized by the US Senate of the first Asian American Senate Page, and there is no widely disseminated information found on Asian American participation in the Senate Page Program. This absence is a clear gap in the Senate records as well as academic literature. Researchers suggest that triangulation, which involves using multiple sources to cross-verify findings, is effective for studying underrepresented groups that are not recognized in historical research (Jick 602; Kipping et al. 305; Noble and Heale 67). Therefore, this study addresses this gap by asking: To what extent can triangulation of archival sources identify early Asian American Senate Pages

who are not formally recognized in institutional records? This study uses triangulation of archival sources as a method to identify the earliest confirmed documented case of an Asian American Senate Page.

Literature Review

Institutional Memory

Studies focusing on historiography and institutional memory show that historical narratives are shaped by not only what is recorded, but equally as much from what is omitted (Trouillot 26). The official memory of an institution is typically shaped by the events and individuals it selectively preserves, which results in unpreserved events and individuals being neglected in official records. This produces an official memory that reflects the priorities of institutions and the contexts to which records were created (Schwartz and Cook 1–19). Because of this, institutional "firsts" are often identified after the fact through research rather than through contemporary official documentation.

The controversy behind the first female Senate Pages reflects this process. In 1971, after debates within the Senate Rules and Administration Committee regarding the controversial appointment of female Senate Pages, Paulette Desell, Ellen McConnell, and Julie Price were sworn in as the "first" female Senate Pages with much media coverage (United States Senate, *Appointment* 3, 16, 29). The timing of their appointment in 1971 also aligned with second-wave feminism and the opening of political institutions to women, so it would have been convenient for Senate institutional narratives to officially recognize the individuals as the "firsts" in official documentation. However, archival work published in 2024 by the Senate Historical Office discovered Emma Madeen, a female Senate Page who served in 1907, demonstrating that the "firsts" identified in 1971 were not truly the earliest female Pages in the US Senate (United States Senate, *Appointment* 3, 16, 29; United States Senate, "When Is a Senate First"). For this situation, the official institutional memory of the US Senate did not change due to a shift in reality, but because new documentation was uncovered. This incident clearly demonstrated that finding institutional "firsts" requires sifting through historical documentation with consideration of narrative biases, and that earlier "firsts" may exist.

Another reason why earlier "firsts" may exist is because demographic information is missing in US Senate institutional memory. In the *Report of the Secretary of the Senate* publications, which can be used as employment records, names and positions of US Senate employees, including Senate Pages, are listed, but the records do not indicate any demographic data including race, gender, and ethnicity (Petersen 2; Jones and Triguero Roura). This brings up the challenge of being unable to use official institutional documentation to identify the demographics of Senate Pages. As a result, researchers rely on external records, such as newspapers, to confirm identities. As seen from the example of the 1971 female Senate Pages, newspaper records can be mistaken and biased due to narrative

biases, making it possible that earlier "firsts" exist. Thus, when milestones for marginalized groups are not documented institutionally, their presence can remain hidden.

Representation and Symbolic Meaning

Political science theories show why identifying "firsts" matters. Pitkin's study distinguishes descriptive representation, where politicians share salient characteristics (such as race, gender, ethnicity) with their constituents, and symbolic representation, where politicians represent the constituents in a way that affects how the group sees its own legitimacy and inclusion (Pitkin 60–92). Mansbridge extends this to institutions by arguing that descriptive representation can improve trust and communication where marginalized groups feel a sense of doubt toward the institutions that serve them (Mansbridge 628).

These studies show the power of historical "firsts." When pioneers are publicly recognized institutionally, it helps develop an inclusive narrative that can affect how people of underrepresented groups relate to those institutions. This recognition can signal that an underrepresented group's presence is legitimate and valued. On the other hand, when pioneers are not recognized or omitted from the official narratives, the participation of an underrepresented group may occur without contributing to symbolic representation in the official narrative of the institution.

Asian Americans' Political Invisibility

Although Asian Americans participate in politics and hold office, their contributions are often omitted in institutional narratives and scholarship. Thus, Asian American representation in politics has been described as a case of "invisibility" (Kim 118). Additionally, Asian Americans are oftentimes left out of discussions about race and representation despite facing discrimination and exclusion (Wu 173). Overall, the political participation of Asian Americans has often been overlooked, which led to a lack of existence in public and institutional memory (Lien et al. 141).

When institutions like the US Senate fail to recognize Asian American "firsts," this invisibility is compounded. Without the institutional documentation, it can blur the line for when Asian Americans entered political institutions, and make it seem as if Asian Americans have only entered political institutions recently, even when their participation stretches back decades. The fact that no officially recognized first Asian American Senate Page exists corroborates this problem: it shows there is a gap in the history of Asian Americans and their contributions with the federal government.

Belonging and Historical Recognition

In addition to representation, studies show that belonging is a key component of inclusion. Belonging is the sense of feeling recognized, valued, and included in institutions. Specifically, it is affected by both the experiences of the individual, and the historical representation the individual's group has in the institution's narratives (Lien et al. 45; Mansbridge 632). Additionally, if a member of the individual's group is in a position of

authority, or if there is an awareness of the history of predecessors of the individual's group, it can strengthen a sense of belonging (Gay 717). Individuals understand their relationship to institutions through historical recognition. When pioneers are acknowledged officially through institutions, it can make it easier for future generations to see themselves belonging in similar roles. Conversely, when there is no recorded institutional history regarding a group in positions within the institution, it can reinforce feelings that the presence of the group is against the norm or unwelcome. Especially in major federal institutions like the Senate, representation in institutional history can be consequential.

Gap in the Literature

In sum, previous research about institutional memory, representation, and political invisibility show that the recognition of "firsts" can affect historical and institutional narratives as well as inclusion and sense of belonging. However, there is no official US Senate documentation or academic research on Asian Americans in the US Senate Page Program, and there is no widely disseminated information regarding the first Asian American Senate Page. This study aims to fill this gap by developing a method to find Senate Page "firsts." Due to the absence of demographic data in official reports and the limitations of using a single source, this study uses triangulation of archival sources and assesses the method by identifying the first Asian American Senate Page.

Methodology

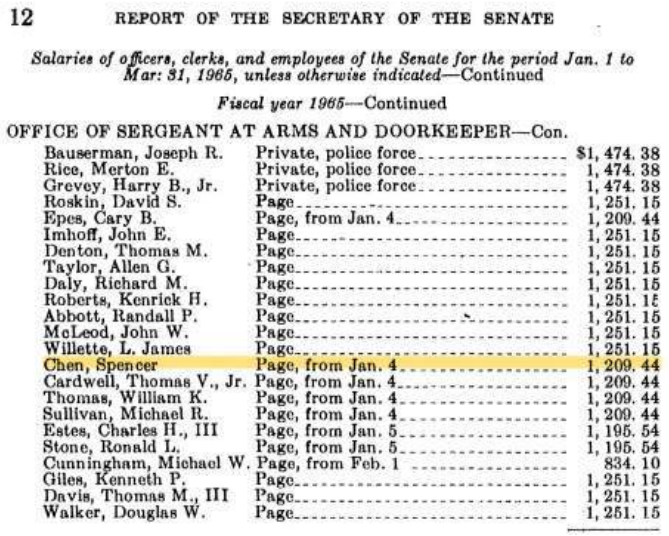
Research Design

To identify the first Asian American Senate Page despite the lack of demographic data in the *Report of the Secretary of the Senate*, this study developed an archival research method based on the triangulation of archival sources. Triangulation involves using multiple sources to cross-verify findings when direct evidence is unavailable or incomplete (Jick 602; Noble and Heale 67). When studying underrepresented groups that are not recognized in historical and qualitative research, triangulation is widely used (Kipping et al. 305).
The study proceeded in three main stages:

1. Identification of potential Asian American Senate Page candidates from the *Report of the Secretary of the Senate* using a Surname-Based Identification (SBI) strategy.
2. Confirmation of participation and refinement of identifying details (e.g., photographic and contextual details, including hometown) through Senate Page yearbooks
3. Contextualization and verification through historical newspapers and other external sources

This triangulation-based process was designed with the intent to maximize credibility through cross-confirmation of findings and produce a method that could be replicated in future studies.

Stage 1: Official Senate Records (see fig. 3)



12 REPORT OF THE SECRETARY OF THE SENATE

Salaries of officers, clerks, and employees of the Senate for the period Jan. 1 to Mar: 31, 1965, unless otherwise indicated—Continued

Fiscal year 1965—Continued

OFFICE OF SERGEANT AT ARMS AND DOORKEEPER—Con.

Bauserman, Joseph R.	Private, police force	$1, 474. 38
Rice, Merton E.	Private, police force	1, 474. 38
Grevey, Harry B., Jr.	Private, police force	1, 474. 38
Roskin, David S.	Page	1, 251. 15
Epes, Cary B.	Page, from Jan. 4	1, 209. 44
Imhoff, John E.	Page	1, 251. 15
Denton, Thomas M.	Page	1, 251. 15
Taylor, Allen G.	Page	1, 251. 15
Daly, Richard M.	Page	1, 251. 15
Roberts, Kenrick H.	Page	1, 251. 15
Abbott, Randall P.	Page	1, 251. 15
McLeod, John W.	Page	1, 251. 15
Willette, L. James	Page	1, 251. 15
Chen, Spencer	Page, from Jan. 4	1, 209. 44
Cardwell, Thomas V., Jr.	Page, from Jan. 4	1, 209. 44
Thomas, William K.	Page, from Jan. 4	1, 209. 44
Sullivan, Michael R.	Page, from Jan. 4	1, 209. 44
Estes, Charles H., III	Page, from Jan. 5	1, 195. 54
Stone, Ronald L.	Page, from Jan. 5	1, 195. 54
Cunningham, Michael W.	Page, from Feb. 1	834. 10
Giles, Kenneth P.	Page	1, 251. 15
Davis, Thomas M., III	Page	1, 251. 15
Walker, Douglas W.	Page	1, 251. 15

Fig. 3. Excerpt listing Spencer Chen as a Senate Page. *Report of the Secretary of the Senate from January 1, 1965, to June 30, 1965*. United States Senate, 1965.

In the first stage, the study utilized the *Report of the Secretary of the Senate*, a semiannual expenditure report of the Senate, including the salary of Senate Pages. Because the *Report of the Secretary of the Senate* has been consistently published since 1823, before the appointment of the first Senate Page, all Senate Pages ever appointed are listed in the reports. For this reason, these reports served as rosters of the Senate Page classes, allowing the identification of the first and last names of Pages as well as what year(s) they served.

Because the *Report of the Secretary of the Senate* does not include race or ethnicity (Petersen 2; Jones and Triguero Roura), the study employed a Surname-Based Identification (SBI) strategy as an initial screening method to identify potential Asian American Senate Pages. SBI is a method used frequently by the U.S. Census Bureau to estimate ethnic populations in official records where there is no demographic data present (Lauderdale and Kestenbaum 285; United States Census Bureau). Surnames that appeared to be potentially of Asian heritage were flagged for further investigation in the next stage. However, it is crucial to note that this stage did not treat the surnames as a single proof of Asian American identity. It was used simply as a way to narrow down the potential candidates for more precise inquiry.

Stage 2: Senate Page Yearbooks (see fig. 4)

Fig. 4. Spencer Chen's yearbook entry. *The Congressional 1965*, Congressional Page School, 1965, p. 36.

The Senate Pages identified as potentially Asian American in the first stage were cross-referenced against the annual Senate Page yearbook. Yearbooks served as a valuable source of verification as they provided photographs and contextual details that helped confirm the identities of candidates identified in the first stage. Furthermore, the yearbooks also include additional information regarding the Senate Page that could refine search inquiries in the third stage. For example, the yearbooks include hometowns/home addresses and party affiliations.

Stage 3: Contemporaneous Media and External Sources (see fig. 5)

Fig. 5. Archival evidence concerning Spencer Chen: **(a)** Spencer Chen with Senators Hiram Fong and Maurine Neuberger, from Tom Stimmel's "Portland High School Youth Makes Washington History," *The Oregon Journal*, 8 Apr. 1965, p. 12; **(b)** Spencer Chen presenting a Chinese painting by his father, Bong Wai Chen, to Senator Everett M. Dirksen, *The Oregon Journal*, 30 Aug. 1965, p. 2; **(c)** Spencer Chen discussing his Senate Page experience in "Bong Wai Chen: Family Oral History," Portland Chinatown Museum, 27 Sept. 2022.

The third stage involved, for each Asian American Senate Page candidate, searching newspaper databases using combinations of the individual's name, hometown, and the term "Senate Page" to find historical newspaper articles confirming the Asian American identity as well as other contextual details such as whether the individual was the first Asian American Senate Page. It is important to note that the Asian American identity of the flagged Senate Pages in the previous stages was confirmed in this stage through written statements found in external sources.

In particular, local newspaper coverage was important for determining whether the flagged Senate Pages were publicly recognized as "firsts" and members of a particular group (e.g. Asian American). Additionally, the study drew on oral history projects, such as an interview with a former Senate Page in a museum collection. As a whole, these sources corroborated the identity of the flagged Senate Pages and added additional context.

Methodological Justification

The triangulation of archival sources was necessary due to the lack of available institutional data and previous studies. Triangulation also allowed for cross-referencing of information found in external sources, compensating for the fact that official Senate records do not document demographic information about Senate Pages. Relying only on names in employment lists would have been methodologically weak and potentially misleading. By combining official records, internal publications, and external media accounts, the study leverages multiple independent lines of evidence to support or rule out potential candidates (Jick 602; Noble and Heale 67). This cross-referencing of different independent sources is based on best practices in previous historical studies on marginalized groups with insufficient established data.

An important strength of this design is that it can be replicated. The three-step process provides a template based on triangulation of archival sources that future studies can apply to different groups or time periods.

Results

Identification of the First Asian American Senate Page

Using this triangulation method, this study found Spencer Chen as the earliest known candidate for the first Asian American Senate Page. This demonstrates that triangulation is a valid method that can identify individuals who were not documented or recorded within official documents. Chen is listed in the 1965 *Report of the Secretary of the Senate* as a Senate Page. By cross-referencing that record with *The Congressional*, the yearbook of Senate Pages during that session, it was determined that he served as a Republican Senate Page. It noted him as being from Portland, Oregon, and that he had been elected class treasurer (Congressional Page School 36).

Further investigations were made on additional sources, including an interview with Chen. More specifically, this interview with Chen was conducted by the Portland Chinatown

Museum for a documentary about him (Portland Chinatown Museum). He talks about his experiences as a Senate Page and remembers being told at the time he was the first Asian American to be a Senate Page. When directly asked if he was “the first” Asian American Senate Page, he said "that's what they told me.”

The newspaper article analysis provided more evidence for Spencer Chen's identification as the first Asian American Senate Page. Three Oregon newspaper articles documented Chen's appointment using racial and historical language to mark his service. A 1965 Oregon Journal article described Chen as "the first Page of Oriental extraction in Senate history" (Stimmel 12). A 1965 Oregon Daily Journal article said he was "the first Oriental to serve as a US Senate Page" (Porter 2). A later 1968 Sunday Oregonian article identified Chen as "the first non-Caucasian Page in the US Senate" (Doyle 137). While it is now widely accepted that the first African American Page in 1869 served as the first known non-Caucasian Senate Page, this fact was not acknowledged at the time. Although the language used by these articles is outdated, the multiple references across separate newspapers provide important corroborative evidence that Chen's service was recognized at the time as a milestone in Senate Page history.

Additional newspaper evidence supported the Asian American identity of Spencer Chen by linking him to Portland's Chinese American culture through his heritage. A Sunday Oregonian article from 1968 referred to Spencer Chen as "son of Bong Wei Chen" and described his father as a previous Portland resident, founder of the Chinese Studio in Portland and artist (Doyle 137). A corresponding obituary for Bong Wai Chen indicated that he was born in Hoy Yen, Canton, China and immigrated to the US in 1937 (“Obituary: Chinese Leader” 4). He later migrated to Portland in 1949. The obituary defined him as artist, journalist, educator and unofficial spokesperson for Portland's Chinese-American Community, referring to his association with Chinese organizations, newspapers (the Oregon Chinese News) and schools (Chinese Language School of the Consolidated Chinese Benevolent Association). These descriptions are evidence of Chen being both a historic senate Page and an active member of a Chinese-American family with established roots within the Chinese community of Portland.

The combination of official records, institutional publications, oral interviews, and local news articles demonstrates very strongly that Chen did serve as a Senate Page in 1965 and was formally recognized as the first person of Asian heritage to do so. Although there is a possibility that an earlier Asian American served as a Page without this being documented, no evidence of another individual was found from this study’s in-depth investigation process.

Subsequent Asian American Senate Pages

After the appointment of Spencer Chen in 1965, there were subsequent appointments of Asian American US Senate Pages. These individuals were flagged during the process of the triangulation of archival sources as Asian American Senate Pages, but were eventually proven to not be the first Asian American US Senate Page through the discovery of an earlier Asian American Senate Page, Spencer Chen.

Winston Lau, appointed in 1972 by Senator Hiram Fong of Hawaii, the first Asian

American senator, was the second Senate Page appointed from Hawaii ("It's Senate Page Lau" 25). He was a Chinese American junior from McKinley High School in Honolulu, Hawaii, Senator Fong's alma mater, and was selected out of 40 applicants ("Paging Mr. Lau" D-7). Senator Fong played an active role in appointing new minorities to the Senate Page Program (see fig. 6), as he previously appointed the first Senate Page from Hawaii, Cliff "Palikumaikalewalani" Needham Ahue, who became the first Native Hawaiian Senate Page in 1971 (Taylor F-2; Osher).

Fig. 6. First appointments by Senator Hiram Fong: **(a)** Cliff Ahue, the first Senate Page appointed from Hawaii, from Lois Taylor's "Page No. 1," *Honolulu Star-Bulletin*, 26 Aug. 1971, p. F-2; **(b)** Senator Hiram Fong with Senate Pages Winston Lou, left, and Cliff N. Ahue, right, photograph by Albert Yamauchi, "The Boys and the Boss," *Honolulu Star-Bulletin*, 25 Dec. 1971, p. 15. Both via *Newspapers.com*.

In addition, another Page from 1972, Mari Iwashita of Cheverly, Maryland (see fig. 7), appointed by Senator Daniel K. Inouye of Hawaii, made history as the first female Asian American Senate Page. Mari Iwashita appeared in many newspapers nationally because she was thought of to be one of the first female Senate Pages regardless of race (AP Wirephoto 11A). Additionally, Mari Iwashita was vocal to the press about the restrictions of female Senate Pages being unable to enter the Senate Democratic Cloakroom, and was portrayed as a controversial "women's lib" activist by reporters (Barr A-24; United Press International D-14). However, with the recent discovery of Emma Madeen, a female Senate Riding Page from 1907, the 1971-1972 "first" female Senate Pages are no longer considered the first female Senate Pages by the Senate Historical Office (United States Senate, "When Is a Senate First"). However, based on the triangulation of archival sources conducted in this study to identify Asian American Senate Pages, no previous female Asian American Senate Pages were found. Therefore, this study deduces that Mari Iwashita is the earliest known female Asian American Senate Page.

Fig. 7. Mari Iwashita, far left, the first female Asian American to serve as a US Senate Page, with fellow female Pages on the US Capitol steps. AP Wirephoto, p. 11A.

Contextualizing 1965

Chen's 1965 experience as a Page was important for two reasons. First, he had achieved what no one else had done at that time. Currently, the only earlier known non-white Senate Page prior to Chen is Andrew Foote Slade, the first known African American Senate Page who served from 1869 to 1881, making Chen the first non-white Senate Page to serve in nearly a century. Ironically, Lawrence Bradford Jr., the first African American Senate Page known at the time who served in 1965 alongside Chen, was mistakenly celebrated by national news networks and Senate institutional records as the first Senate Page of color, even though Chen had been appointed two months earlier and was already serving on the Senate floor. This shows that even though Chen was rendered invisible, he pioneered the racial integration of the Senate Page Program during the Civil Rights era. Second, the time in which Chen served as a Page (mid-1960s) was a time when the United States experienced great changes socially and politically. There were many key events during this time such as the Civil Rights Movement, the passage of major civil rights legislation, and growing awareness of racial equality and representation. In light of these developments, it was symbolic that there would be a young Asian American person, working daily in a very prominent government building, representing some of the first racial diversity in the Federal Government.

The details provided by the 1965 yearbook from the Congressional Page School further support that Chen was not simply an observer. He participated at a level that would indicate to others that they could rely upon him for a leadership role as indicated by the fact that his fellow classmates elected him as class treasurer. This suggests that he had become accepted into this group which lends additional evidence that it was not just symbolic that he was there, but it was through his involvement in day-to-day functions of the program that he made a difference.

The use of language in the newspapers of the 1960s provide some historical context for how Chen's identity was perceived at that time. That he was described as being of “Oriental descent" is an example of the ways in which Chen's race/ethnicity were acknowledged by

others, while also reflecting the racial terms used during that period. Although such language would be viewed as outdated and often problematic today, it does indicate that Chen's race/ethnicity was meaningful to Americans at that time. Thus, this also contributed to why his selection as a Senate Page was newsworthy. Together, these examples illustrate that Chen's experience as a Senate Page had great significance in 1965; however, it has been omitted from official institutional history.

Discussion

Institutional Limitations and the Formation of "Firsts"

This study shows how institutional "firsts" may be developed with incomplete documentation. Since the Senate Page Program did not collect data regarding race/ethnicity, it was impossible to simply look at a database to find out the participants' race/ethnicity. The researcher therefore had to infer Senate Pages' racial identity by looking at multiple sources. Thus, such use of indirect methods is indicative of the limitations that the documentation process imposes upon researchers who attempt to reconstruct history.

This instance further illustrates how institutional history is a selective or incomplete account (Trouillot 26). That is to say, the institutional history of the US Senate's Page Program includes some major events in the program's history (e.g., the acceptance of both women and African American Pages), but it excludes many other significant events. Such omissions do not mean that those events never happened (Moss and Thomas 11); rather, they signal what was documented and later become part of a collective memory. The discovery of female Pages from 1907 after decades of recognizing 1971 as "the first" is one example. The first Asian Senate Page, which had long remained overlooked and neglected in the historical record, could be another example.

The fact that Chen may have been present in local news, but absent from formal histories indicates that institutional recognition does not automatically follow. The "firsts" produced through institutional means are therefore, in many ways, a product of continuous academic inquiry, reinterpretation, and debate on new information as it is discovered. These results also show how triangulating data (as described above) provides a strong methodology for identifying likely or well supported cases of possible firsts, however, it will never provide an absolute history of firsts if documentation is limited.

Visibility Versus Recognition

The difference between "visibility" and "recognition" is what emerges clearly when we look at Chen's situation. The term "visibility," refers to the presence of representation. He served as a Senate Page, and people may have seen him because there was media reporting on his experience at the time. "Recognition," however, would refer to the official acknowledgment and inclusion of that representation within an institution's historical account(s), and/or public memory.

Visibility existed for Chen through local newspaper reports at the time and his own

community knowing about his accomplishments. However, this visibility did not equate to institutional recognition; no record exists with regard to the US Senate Page Program formally recording or celebrating him as their first Asian American Senate Page. Therefore, although he made significant contributions toward the diversification of the program, his achievements have never been included in general historical records of such diversifications (although many of these have been recorded and are celebrated).

This pattern may also apply to many other kinds of selective memories in history (Schwartz and Cook 1–19). For instance, institutions frequently focus on different forms of diversity (e.g., promoting women in the military), however, at the same time, they tend to ignore inclusions with respect to other types of diversity. Within the Senate Page Program, significant milestones related to African Americans and women have received documentation and celebration. On the other hand, milestone events associated with Asian Americans remain unrecognized. While this may not be evidence of an institutional intent to intentionally avoid recognizing Asian Americans, it illustrates how institutional priorities can shape which "firsts" are remembered.

In addition, time is another critical aspect to consider. In general, visibility occurs for a short duration: a brief newspaper article generates interest for only a short period of time before it disappears into oblivion. However, recognition requires that an institution formally acknowledges something/someone over time. Chen may have been visible during 1965; however, due to the fact he never received formal recognition for being the first Asian American Senate Page, his visibility declined. Therefore, his pioneering work failed to become part of how the subsequent generations of Senate Pages would tell the story about the program.

Representation and Sense of Belonging

Chen's inclusion into the Senate Page Program is a strong refutation for the belief that Asian Americans had little to no influence within governmental institutions until the recent past. In fact, his time as a Senate Page in 1965 indicates that Asian Americans were in the US Senate many years before they would be widely recognized for such political engagement.

The fact that Chen was the first Asian American to serve as a Senate Page from the standpoint of representation theory is a further addition to the history of descriptive representation in federal institutions. While his presence as a Page had no quantitative effect on the demographics of the Senate, he was placed in a position where an Asian American could be seen by others while working at the center of Legislative Activity. The presence of individuals such as Chen can have symbolic effects as previous studies have suggested (Mansbridge 628; Pitkin 60–92). The knowledge that there has been an Asian American teenager who walked down the corridors of the Senate as a Page can shape how later generations of Asian American students envision themselves within US politics (Wolbrecht and Campbell 921). This research does not directly measure "sense of belonging," however, it emphasizes that historical recognition can shape a sense of belonging by signaling inclusion. When institutions recognize the contributions of historically underrepresented groups, they signal that these groups are a component of the institution's narrative and provide a sense of inclusion (Rocha et al. 890).

When institutions fail to recognize Asian American "firsts," this reinforces feelings of being marginalized or being perpetual outsiders (Kim 106–07).

Limitations

While the current study produced meaningful results strengthened by the triangulation of archival sources, it also has several limitations. First, the initial screening method for identifying candidates involved using the Pages' last names. A major drawback of this technique is the likelihood of obtaining false positives (last names that sound Asian but are not associated with individuals who are actually Asian) and false negatives (Asian Americans whose last names do not signal that identity). While these data were used simply to identify candidates for additional examination, they implied incompleteness in the search procedure. Secondly, given that the *Report of the Secretary of the Senate* and related documentation do not contain racial/ethnic identifiers, it cannot be firmly concluded whether or not any Asian Americans served as a Senate Page prior to 1965. Even though the study suggests that Spencer Chen is likely the first documented Asian American Senate Page based upon strong evidence, earlier unidentified or undocumented cases could potentially exist. Newspaper articles and an oral interview were used to verify the findings, but both possess inherent flaws. They can offer rich detail and perspectives on events; however, they may contain biased information, and personal testimonies may be strongly influenced by one's own memories and experiences. These limitations may affect the reliability of the findings.

Implications and Future Research

The findings of this study offer implications for both institutional practices and scholarly work. On the institutional side, the study suggests the need for a critical evaluation of historical records associated with an institution. Through the example of Spencer Chen's contribution as a "first," we see that contributions may be overlooked due to voids in documentation. In order to further document such "firsts," systematic evaluations of existing archives should occur on a regular basis. These evaluations should specifically seek out new documentation regarding historically underrepresented groups and provide additional insight into a more inclusive representation of the history of those who were involved within an institution.

For researchers, this study shows that the triangulation of archival sources can provide effective results where official documents are either limited or missing. Future research may consider using a similar methodology to identify the "first" or earlier representation of underrepresented groups. Another area of investigation could include looking at how recognition of the past affects present-day experiences. Former and current Pages could be interviewed or surveyed to determine if knowing about previous participants, such as Chen, enhances their feelings of belonging, increases their ambitions politically, or changes their views towards institutions. Researchers could also look into how institutions record and publicly report on diversity, to see if there are identifiable patterns concerning which groups tend to be recognized and which continue to be unreported. Further research on the overlooked "firsts" may also provide a broader perspective about the political development in the United States. By filling in

gaps within the current literature and correcting misconceptions regarding participation in governmental institutions, it can support creating narratives that more accurately represent the diversity of the United States.

Conclusion

The current study shows that the triangulation of archival sources is a useful method for identifying early United States Senate Pages absent from institutional records, as shown from the case of Spencer Chen. The identification of institutional "firsts" is a complex process limited by the availability of archival documentation and institutional memory. When official records omit demographic information, additional external sources are often necessary to capture the full diversity of political institutions. Utilizing the triangulation of archival sources is useful for critical engagement of external sources and available institutional data.

The identification of Spencer Chen also shows the gap between visibility and recognition. Even though his appointment and service as a Senate Page was recognized in a local newspaper, he was forgotten in the US Senate's institutional records. This omission demonstrates Asian American political invisibility as well as how institutional narratives can exclude certain groups. In terms of representation and belonging, recovering Chen's story matters because it adds to the broader narrative of diversity in the US Senate Page Program and challenges the notion that Asian Americans were absent from elite federal institutions until recently.

Documenting Spencer Chen's service in the US Senate can inform how later generations of Asian American Senate Pages understand their relationship to the Senate Page Program and the Senate. Ultimately, this study demonstrates how identifying ways to uncover lost histories will help to shift an institution's historical story in a more inclusive way. By expanding who is recognized and remembered, institutions can widen narratives on the diversity of individuals who have contributed to their development and foster a sense of belonging among historically marginalized groups.

Works Cited

Amer, M. L. Pages of the United States Congress. U.S. Government Printing Office, 1981.

AP Wirephoto. “Pages on Record.” Photograph caption. Pueblo Chieftain, 22 Nov. 1971, p. 11A. Newspapers.com.

Appalachian Regional Commission. Abigail Spanberger. Appalachian Regional Commission, https://www.arc.gov/staff/abigail-spanberger/. Accessed 23 July 2026.

Barr, Malcolm. “Washington and Hawaii.” Honolulu Star-Bulletin, 22 Nov. 1971, p. A-24. Newspapers.com.

Budiman, Abby, and Neil G. Ruiz. “Asian Americans Are the Fastest-Growing Racial or Ethnic Group in the U.S.” Pew Research Center, 9 Apr. 2021, https://www.pewresearch.org/short-reads/2021/04/09/asian-americans-are-the-fastest-growing-racial-or-ethnic-group-in-the-u-s/.

Clayton, Amanda, et al. “All Male Panels? Representation and Democratic Legitimacy.” American Journal of Political Science, vol. 63, no. 1, 2019, pp. 113–29. https://doi.org/10.1111/ajps.12391.

Congressional Page School. The Congressional 1965. 1965.

Dietsch, Kevin. U.S. Sen. Mike Lee Participates in a Senate Judiciary Committee Hearing. 13 May 2025. Getty Images. Wisconsin Examiner, 17 June 2025, https://wisconsinexaminer.com/2025/06/17/sen-mike-lee-takes-down-x-posts-after-widespread-criticism/. Accessed 23 July 2026.

Doyle, J. “Year of the Monkey.” The Sunday Oregonian, 28 Jan. 1968, p. 137.

Gay, Claudine. “Spirals of Trust? The Effect of Descriptive Representation on the Relationship between Citizens and Their Government.” American Journal of Political Science, vol. 46, no. 4, 2002, pp. 717–32.

“It’s Senate Page Lau.” The Honolulu Advertiser, 8 Dec. 1971, sec. B, p. 25. Newspapers.com.

Jantzen, Franz. Associate Justice Neil M. Gorsuch. 25 May 2017. Collection of the Supreme Court of the United States. Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Associate_Justice_Neil_Gorsuch_Official_Portrait.jpg. Accessed 23 July 2026.

Jick, Todd D. “Mixing Qualitative and Quantitative Methods: Triangulation in Action.” Administrative Science Quarterly, vol. 24, no. 4, 1979, pp. 602–11.

Jones, J. R., and M. Triguero Roura. “Inequality on Top of the Hill: Race, Pay, and Representation among Congressional Staff Members.” Socius: Sociological Research for a Dynamic World, vol. 11, 2025, article 23780231251397012. https://doi.org/10.1177/23780231251397012.

Kim, C. J. “The Racial Triangulation of Asian Americans.” Politics & Society, vol. 27, no. 1, 1999, pp. 105–38. https://doi.org/10.1177/0032329299027001005.

Kipping, Matthias, et al. “Analyzing and Interpreting Historical Sources: A Basic Methodology.” Organizations in Time: History, Theory, Methods, edited by Marcelo Bucheli and R. Daniel Wadhwani, Oxford UP, 2014, pp. 305–29.

Lauderdale, Diane S., and Bert Kestenbaum. “Asian American Ethnic Identification by Surname.” Population Research and Policy Review, vol. 19, no. 3, 2000, pp. 283–300.
Lien, Pei-te, et al. The Politics of Asian Americans: Diversity and Community. Routledge, 2004.
Mansbridge, Jane. “Should Blacks Represent Blacks and Women Represent Women? A Contingent ‘Yes.’” The Journal of Politics, vol. 61, no. 3, 1999, pp. 628–57.
Moss, Michael, and David Thomas. “Theorising the Silences.” Archival Silences: Missing, Lost and Uncreated Archives, edited by Michael Moss and David Thomas, Routledge, 2021, pp. 10–25.
Noble, Helen, and Roberta Heale. “Triangulation in Research, with Examples.” Evidence-Based Nursing, vol. 22, no. 3, 2019, pp. 67–68.
“Obituary: Chinese Leader Here, Bong Wai Chen, Dies in Oakland.” The Oregon Daily Journal, 16 Feb. 1968, p. 4.
“Paging Mr. Lau -- Washington Calling.” Honolulu Star-Bulletin, 8 Dec. 1971, p. D-7. Newspapers.com.
Petersen, R. Eric. Staff Pay, Selected Positions in Senators’ Offices, FY2001–FY2023. CRS Report no. R44324, Congressional Research Service, 2024.
Photograph of Spencer Chen presenting a Chinese painting by Bong Wai Chen to Senator Everett M. Dirksen. The Oregon Journal, 30 Aug. 1965, p. 2. NewsBank. Accessed 6 June 2026.
Photograph of Spiro Agnew. Encyclopaedia Britannica, https://www.britannica.com/biography/Spiro-Agnew. Accessed 23 July 2026.
Pitkin, Hanna Fenichel. The Concept of Representation. U of California P, 1967.
Porter, D. “Proud Father Sends Art to U.S. Officials.” The Oregon Daily Journal, 8 May 1965, p. 2.
Portland Chinatown Museum. “Bong Wai Chen: Family Oral History.” Vimeo, 27 Sept. 2022, https://vimeo.com/830451776.
Rocha, Rene, et al. “Race and Turnout: Does Descriptive Representation in State Legislatures Increase Minority Voting?” Political Research Quarterly, vol. 63, no. 4, 2010, pp. 890–907.
Schwartz, Joan M., and Terry Cook. “Archives, Records, and Power: The Making of Modern Memory.” Archival Science, vol. 2, 2002, pp. 1–19.
Stimmel, Tom. “Portland High School Youth Makes Washington History.” Oregon Journal, 8 Apr. 1965, p. 12. NewsBank.
Taylor, Lois. “Page No. 1.” Honolulu Star-Bulletin, 26 Aug. 1971, p. F-2. Newspapers.com.
Trouillot, Michel-Rolph. Silencing the Past: Power and the Production of History. Beacon Press, 1995.
United Press International. “Inouye Faults Media over Girl Page.” The Honolulu Advertiser, 24 Nov. 1971, p. D-14. Newspapers.com.
United States Census Bureau. Frequently Occurring Surnames in the 2010 Census. U.S. Department of Commerce, 2016, https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf.

United States Senate. “About Senate Pages: Historical Overview.” United States Senate, https://www.senate.gov/about/officers-staff/Pages/overview.htm. Accessed 25 June 2026.
United States Senate. Appointment of Female Senate Pages: Hearing Before the Ad Hoc Subcommittee to Consider the Appointment of Female Pages of the Committee on Rules and Administration. U.S. Government Printing Office, 1971.
United States Senate. “When Is a Senate First Truly a First?” United States Senate, https://www.senate.gov/artandhistory/senate-stories/when-is-a-senate-first-truly-a-first.htm. Accessed 25 June 2026.
United States Senate Page Program. “History.” United States Senate Page Program, https://pageprogram.senate.gov/history/. Accessed 25 June 2026.
United States Senate Historical Office. First Female Senate Pages. United States Senate, 2014, https://www.senate.gov/about/resources/pdf/first-female-Pages.pdf.
United States Senate Historical Office. Michael F. Bennet. United States Senate, https://www.senate.gov/artandhistory/history/common/image/Bennet_Michael.htm. Accessed 23 July 2026.
“U.S. Senate.” C-SPAN, C-SPAN2, www.c-span.org/congress/?chamber=senate. Accessed 14 Feb. 2026.
Wolbrecht, Christina, and David E. Campbell. “Leading by Example: Female Members of Parliament as Political Role Models.” American Journal of Political Science, vol. 51, no. 4, 2007, pp. 921–39.
Wong, J., et al. “Making Visible: Political Participation.” Asian American Political Participation: Emerging Constituents and Their Political Identities, Russell Sage Foundation, 2011.
Wu, Frank H. Yellow: Race in America Beyond Black and White. Basic Books, 2002.
Yamauchi, Albert. “The Boys and the Boss.” Honolulu Star-Bulletin, 25 Dec. 1971, p. 15. Newspapers.com, www.newspapers.com/image/271446093/. Accessed 18 June 2026.

microRNAs in Alzheimer's Disease: Amyloid -β and Tau Regulation and Their Usage as Biomarkers and Therapeutic Targets By Khurshid Jimmy Hormazdiyar

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disease associated with cognitive impairment, anterograde amnesia, and the presence of amyloid-βeta (Aβ) plaques along with tau -related neurofibrillary tangles in the cerebral parenchyma. Although numerous studies have explored the molecular mechanisms underlying disease onset and progression, the exact mechanisms remain incompletely elucidated. In recent years, microRNAs (miRNAs) have become key components of AD pathogenesis, as small non-coding RNA molecules that regulate gene expression post-transcriptionally. In this review, the role of miRNAs in AD has been critically examined, with particular attention to their contributions to Aβ production and clearance, as well as to tau hyperphosphorylation. Particular miRNAs, miR 29a/b, miR-298, miR-125 b, and miR-132/212 cluster, are mentioned in terms of their ability to regulate the genes of APP, BACE1, and tau-associated kinases. The dysregulation of these miRNAs also leads to increased Aβ accumulation, impaired clearance, increased neuroinflammation, and abnormal tau modification, all of which culminate in impaired neuronal processes and disease progression. Besides their functional roles, miRNAs are also considered non-invasive diagnostic biomarkers due to their stability in biofluids such as plasma and cerebrospinal fluid. New therapeutic approaches, such as miRNA mimics and miRNA inhibitors, are discussed for their ability to simultaneously target multiple disease pathways. This review highlights the utmost importance of miRNAs in AD and emphasizes their potential to both advance the detection of the disease in its initial phases and serve as targets for targeted therapeutic interventions.

1. Introduction

Alzheimer's disease (AD) is a progressive and irreversible brain disease that gradually destroys memory, mental abilities, and the ability to perform even daily activities. It is the leading cause of dementia and primarily occurs in older adults, but there are cases of early-onset as well. The population with AD is expanding across the world, thus becoming a serious public health issue. In AD, symptoms typically start with minimal memory loss and confusion but increase to severe cognitive impairment, changes in behavior, and total dependence on caregivers (Vernet et al.).

Two significant changes within the brain characterize the illness. One is the accumulation of amyloid-βeta (Aβ), which forms sticky plaques between neurons (Selkoe & Hardy, 2016). The plaques are poisonous and may disrupt the way neurons communicate with one another. Second, there is an accumulation of tau protein within neurons **(**Kavanagh et al). Under normal circumstances, tau helps maintain the stability of the cell's transport system. However, in AD, tau becomes aberrant, forming twisted angles that clog this inner transport system. As plaques and tangles progress throughout the brain, neurons begin to die, leading to brain tissue shrinkage and a loss of function (Kavanagh et al.).

Despite extensive research, the precise etiology of AD remains unknown. However, increasing evidence indicates that all factors, including genetic, environmental, and lifestyle aspects, contribute to the disease. In the last few years, one such aspect that has attracted considerable interest is the involvement of microRNAs (miRNAs) in the pathogenesis and progression of AD.

miRNAs are extremely small RNA molecules that do not encode proteins (He and Hannon). Instead, they function as regulators by turning other genes on or off. They do this by binding to messenger RNA (mRNA), which encodes instructions for making proteins. When a miRNA binds to an mRNA, it can silence translation of the mRNA into a protein or lead to its destruction (He and Hannon). This process is significant because it allows the body to regulate which proteins are produced, when they are produced, and how much of them are produced.

Although miRNAs are tiny, they play a critical role in bodily functions. They regulate crucial processes such as cell growth, development, immune responses, stress responses, and most notably, brain activity. In the brain, miRNAs play roles in neuronal development, intercellular signaling (synaptic function), memory acquisition, and protection against injury. As a result, any disturbance in miRNA equilibrium will lead to brain dysfunction. Research has established that imbalances in miRNA levels are associated with most diseases, including Parkinson's disease, Huntington's disease, and AD (Li et al.).

It has been shown that certain miRNAs become more active (upregulated) or less active (downregulated) in AD. This can influence the production of proteins such as Aβ and tau, leading to their toxic accumulation in the brain. For instance, certain miRNAs reduce the expression of enzymes responsible for Aβ production, so when these miRNAs are low or absent, more Aβ is synthesized. Other miRNAs may regulate tau protein stability; thus, their malfunction can cause tangles to form. These observations suggest that miRNAs may play a major role in the initiation and early-stage progression of AD.

Since miRNAs are non-volatile and can be detected in blood, cerebrospinal fluid, and even saliva, they are currently being investigated as potential biomarkers for AD. This would imply that physicians could identify early predictors of AD using basic blood tests by measuring specific miRNA levels. This would enable earlier diagnosis of the disease, before significant symptoms set in, when patients are more likely to be treated.

In addition, researchers are investigating the possibility of miRNA therapies. These include employing miRNA mimics to enhance beneficial miRNAs or miRNA inhibitors to prevent detrimental ones. It is conceivable that modulating the levels of certain miRNAs may decrease Aβ production or prevent tau tangle formation. Although these drugs are still in the experimental stage, they hold great promise for the future of Alzheimer's treatment.

This review aims to discuss the growing body of literature on the interaction between miRNAs and AD. It will focus on how miRNAs influence the formation and accumulation of Aβ and tau proteins. This review will also identify miRNAs that are altered in individuals with AD and assess their roles as diagnostic biomarkers and therapeutic targets. By learning how miRNAs

influence brain health and contribute to the onset of AD, we could develop improved methods for detecting, treating, or even preventing this difficult and prevalent disease.

2. The Overview of miRNAs

2.1 The structure of miRNA, how they are made, and how they operate.

miRNAs are small, non-coding RNA molecules, approximately 20-24 nucleotides in length, and play an essential role in gene control (Bartel). Even though they do not code for proteins directly, miRNAs control gene expression by binding to complementary sites on messenger RNAs (mRNAs), leading to either mRNA degradation or translational repression. Through this pathway, they can control protein production, as required for the maintenance of normal cellular function.

The synthesis of miRNAs begins in the nucleus, where they are transcribed from DNA as long primary transcripts (pri-miRNAs) by RNA polymerase II or III (Lee et al.). These pri-miRNAs contain hairpin structures, are identified and processed by a complex consisting of the enzyme Drosha and the protein DGCR8 into shorter precursor forms called pre-miRNAs. These ~70-nucleotide-long pre-miRNAs are subsequently exported from the nucleus to the cytoplasm through a protein called Exportin-5.

In the cytoplasm, pre-miRNAs are again processed by another enzyme, Dicer, which cleaves them into ~22-nucleotide-long double-stranded mature miRNA duplex. One of the strands of the duplex, referred to as the guide strand, is loaded into the RNA-induced silencing complex (RISC), whereas the other strand, referred to as the passenger strand, is typically degraded. The guide strand of the RISC complex then guides the complex to its target mRNA, resulting in gene silencing by either cleaving the mRNA or blocking its translation, depending on the level of sequence complementarity.

The regulation of miRNAs is strict at various levels, i.e, transcriptional regulation, processing efficacy, and stability. Various factors, including stress, hormones, signaling molecules, and epigenetic changes, can affect miRNA expression. Aberrations in any component of miRNA biogenesis or regulation can lead to abnormally elevated or reduced miRNA levels, which can cause a variety of diseases, including cancer, neurodevelopmental disorders, and neurodegenerative disorders such as AD.

2.2 miRNAs' roles in cellular maintenance

The main role of miRNAs is post-transcriptional regulation of gene expression. They bind to target mRNAs, thus controlling the timing, location, and quantity of protein production. Since they can regulate multiple genes simultaneously, a single miRNA can regulate gene networks and influence various biological processes.

Among the most researched cellular processes controlled by miRNAs are cell differentiation, metabolism, immune responses, and stress adaptation. For example, miR-21 has been found to enhance cell survival and proliferation, whereas miR-34a facilitates apoptosis. In

this way, miRNAs maintain cellular homeostasis and prevent uncontrolled cell growth or cell death under normal conditions.

Another essential feature of miRNAs is tissue-specific expression. Distinct organs and tissues have specific miRNA expression patterns that match their functional demands (Lagos-Quintana et al.). This is especially true in the central nervous system, where miRNAs play important roles in brain development, synaptic plasticity, neuronal differentiation, and cognitive processes such as learning and memory. miR-124, for instance, is highly expressed in the brain and modulates neurogenesis and neuronal identity, whereas miR-132 functions during synaptic plasticity and memory formation.

The brain is perhaps the most miRNA-decorated organ, and disturbance in these brain-specific miRNAs may result in neurological disorders. miRNA dysregulation has been implicated in schizophrenia, autism spectrum disorders, Parkinson's disease, and AD (Kosik). MiRNA profile changes in these diseases indicate that miRNAs not only regulate brain activity but also serve as potential markers of neurological disorders.

2.3 The involvement of miRNA in the development of Alzheimer's Disease.

In the AD, miRNAs can influence APP processing, tau phosphorylation, synaptic function, and neuroinflammation, all of which are hallmarks of the disease (Hebert & De Strooper). Their ability to modulate several aspects of the disease makes them potential therapeutic targets and useful tools for early diagnosis. Numerous studies have shown that many miRNAs are altered in AD and that these changes affect processes such as Aβ aggregation, tau hyperphosphorylation, neuroinflammation, and synaptic disruption (Figure 1).

In summary, miRNAs are potent regulators of gene expression with far-reaching functions across diverse biological systems. Their generation is a highly coordinated multi-stop process, and their deregulation is associated with profound pathological effects. Within the brain, their tissue-specific functions are especially essential for neuronal function, and their implication in AD etiology underscores their role in both disease understanding and the treatment of neurodegenerative diseases.

3. The Overview of AD

3.1 History of AD

AD was originally described in 1906 by Dr. Alois Alzheimer, who autopsied the brain of a patient with marked memory loss, confusion, and bizarre behavior (Maurer et al.). Post-mortem examination revealed unusual clumps and twisted fibers in her brain, now referred to as amyloid plaques and neurofibrillary tangles, respectively. These findings provided the basis for appreciating AD as a specific neurodegenerative disorder, and the disease was subsequently named after him.

3.2 Pathophysiology of AD

AD is a neurodegenerative disease characterized by the progressive degeneration of neurons, synaptic contacts, and brain mass, particularly in memory- and cognition-related brain regions (Selkoe & Hardy). Gradual cognitive decline over time, accompanied by memory loss, language difficulties, impaired problem-solving, and behavioral changes, occurs in patients with AD.

Two hallmarks characterized the AD pathology: tau neurofibrillary tangles and Aβ plaques (Querfurth & LaFerla). These are responsible for widespread damage to brain tissue, interrupting neuronal signaling, and leading to cell death. As neuron loss continues, brain shrinkage becomes apparent, suggesting that the individual gradually loses the ability to live independently.

3.3 Amyloid hypothesis

The amyloid hypothesis suggests that AD starts with the buildup of Aβ peptides in the brain (Hardy & Selkoe). These peptides are produced from amyloid precursor protein (APP) upon cleavage by beta- and gamma-secretase enzymes. In a normal brain, Aβ is removed, but in AD, it clumps, creating sticky plaques between neurons. These amyloid plaques disrupt synaptic communication, causing inflammatory responses, and lead to neuronal damage and death. Huge research efforts to limit Aβ deposition via beta-secretase inhibitors and immunotherapies have yielded inconsistent clinical outcomes (Huang et al.). Nonetheless, recent evidence suggests that reducing soluble oligomeric Aβ species could yield better outcomes and slower cognitive deterioration.

3.4 Tau hypothesis

The tau hypothesis emphasizes changes in the tau protein, which naturally helps stabilize microtubules, structural components of neurons (Iqbal et al.). In AD, tau is hyperphosphorylated, acquiring an excess of phosphate groups, which causes neurofilament separation from microtubules. This leads to the deposition of twisted fibers, known as neurofibrillary tangles, within neurons. The tangles interfere with the cell's internal transport system, hindering the passage of vital proteins and nutrients and leading to cell death. Notably, the extent and tightness of tau tangles correlate more strongly with clinical severity and cognitive impairment in AD patients than do amyloid plaques, underscoring the significance of tau in the disease.

4. Role of miRNAs in AD

miRNAs are small, endogenous, non-coding RNA molecules (~22 nucleotides) that modulate gene expression post-transcriptionally. Through binding to complementary sequences on mRNAs, they repress translation or trigger mRNA degradation (Bartel). In the AD, such molecules are under scrutiny because their regulatory influence is thought to extend to genes involved in Aβ deposition and tau protein hyperphosphorylation, and their potential reversibility positions them as both biomarkers and therapeutic targets.

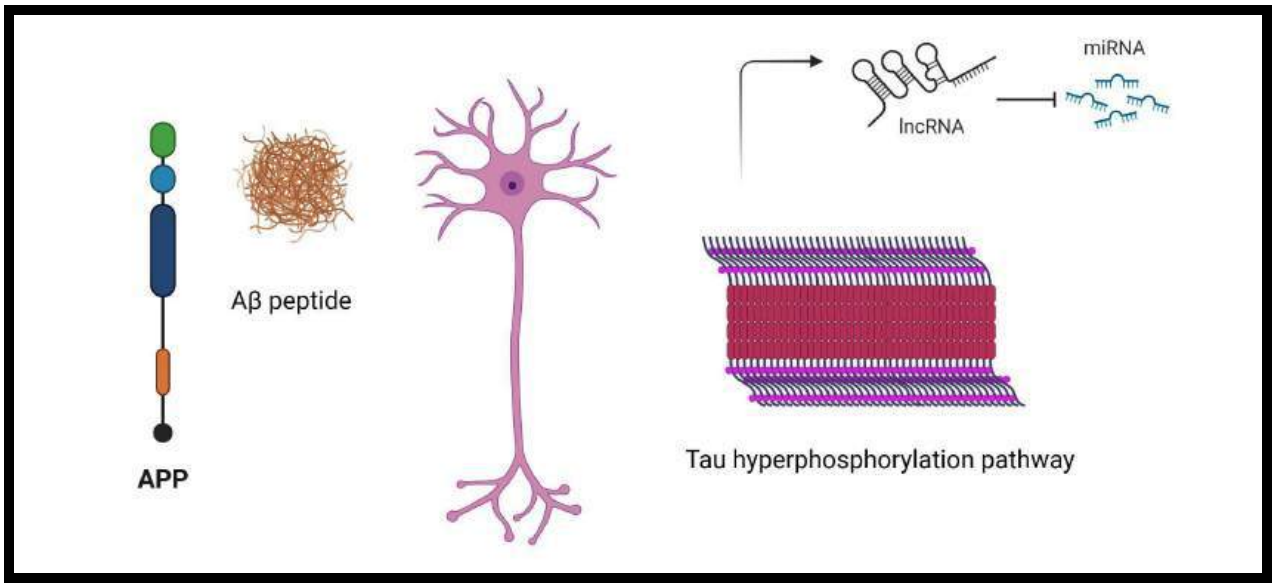


Figure 1. Overview of miRNA involvement in Alzheimer's disease. Dysregulated miRNAs affect the accumulation of amyloid-β (Aβ), hyperphosphorylated tau (pTau), neuroinflammation, synaptic dysfunction, and neuronal death in patients with Alzheimer's disease.

Regulation of Aβ synthesis is mediated by specific miRNAs that negatively regulate the expression of key proteins involved in Aβ production (Delay et al.). A significant reduction in miR-29a/b-1 levels was observed in individuals diagnosed with AD, and this miRNA normally represses BACE1 mRNA, a critical enzyme required for Aβ production (Bettens et al.). A decrease in BACE1 mRNA levels increases BACE1 expression and, subsequently, Aβ. miR-298 also regulates both APP and BACE1, and the level of miR-298 is also decreased in patients with AD, which is correlated with higher levels of Aβ (Baza et al.). Therefore, the loss of these protective miRNAs may be an early molecular alteration that increases the risk of cognitive deterioration and synaptic dysfunction.

AD is characterized by an inability to clear the Aβ peptide, which is due, in part, to the activities of some miRNAs. Both miR-146a and miR-125b are elevated in AD and inhibit the functions of several important genes that regulate microglial phagocytosis, e.g., TREM2 and IRAK1 (Banzhaf-Strathmann et al.). Therefore, the upregulation of these two miRNAs causes a loss of microglial ability to clear Aβ from the brain. Conversely, miR-34a and miR-155 are both increased by Aβ and promote the progression of neuroinflammation and oxidative stress. These factors further disrupt microglial functions in clearing Aβ.

Low levels of miR-339-5p correlate with increased BACE1 levels and decreased glial cell function, indirectly leading to Aβ buildup in the brain (Woods et al.). Dysregulation of these miRNAs contributes to extracellular accumulation of Aβ, plaque formation, neuronal damage, and synaptic dysfunction. Specific miRNAs are involved in regulating the production and clearance of Aβ through the APP, the β-secretase 1 (BACE1) enzymes that are responsible for Aβ production, and the microglial pathways that help to clear Aβ from the brain (Figure 2).

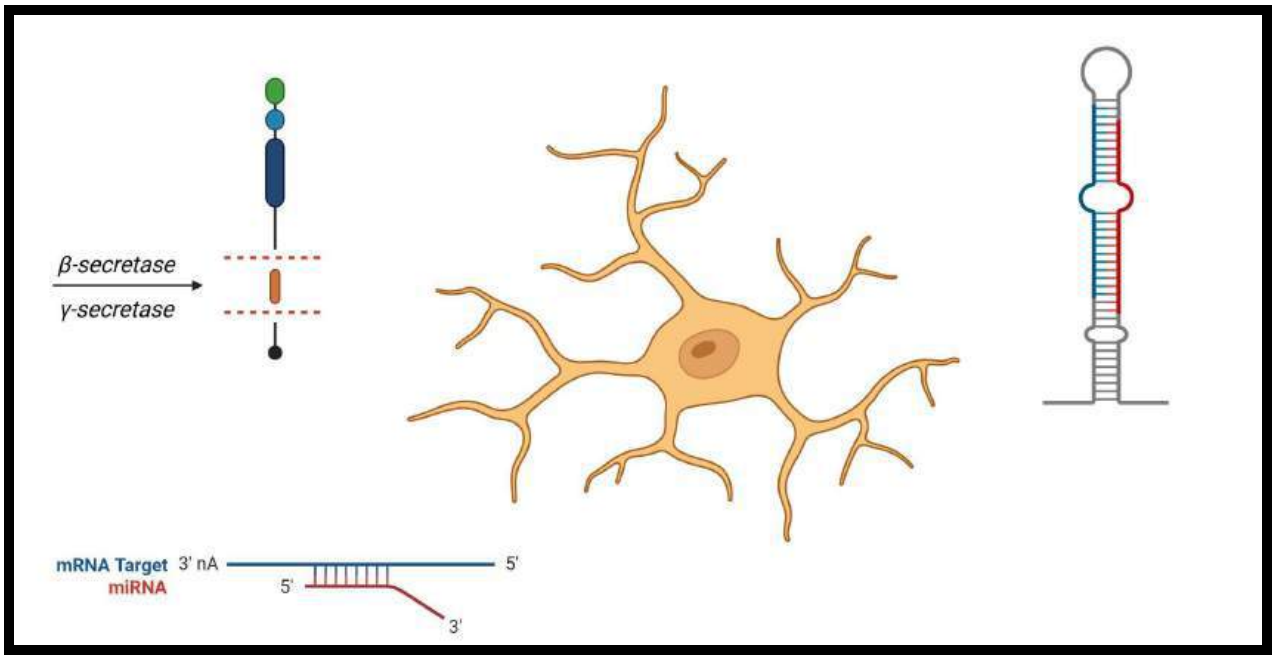


Figure 2. miRNAs regulating Aβ production and clearance. Specific miRNAs regulate APP, BACE1, and microglial clearance pathways, thereby influencing amyloid-β (APP) production, plaque deposition, and disease progression in patients with Alzheimer's disease.

According to the tau hypothesis, hyperphosphorylation of the tau protein leads to the dysregulation of normal neuronal processes and the degeneration of these neurons due to tau neurofibrillary tangles (Long et al.). The dysregulation of various miRNAs has been shown to affect this process. For example, increased expression of miR-125b has been shown to promote hyperphosphorylation of tau by upregulating tau kinases (CDK5, ERK1/2) and downregulating tau phosphatases, resulting in neuronal toxicity (Banzhaf-Strathmann et al., "MicroRNA-125b Induces Tau Hyperphosphorylation and Cognitive Deficits in Alzheimer's Disease"). Conversely, decreased levels of miR-132/212 in patients with AD have been shown to enhance tau phosphorylation, thereby impairing synaptic plasticity. Other miRNAs, such as miR-504-3p and miR-26b, are additional regulators of tau-related kinases, suggesting that aberrant regulation of miRNA contributes to tau-mediated neurodegeneration.

The use of miRNAs targeting both the Aβ and tau pathways is particularly significant in AD research, as these miRNAs simultaneously interact with and modulate multiple core mechanisms of the disease. miR-298 has been shown to regulate the expression of APP, BACE1, and some tau-associated kinases, thereby decreasing amyloid levels and reducing tau pathology (Long et al., 2012). When researchers experimentally overexpressed miR-298 in a mouse model of AD, they demonstrated substantial improvements in memory and cognitive abilities, underscoring the neuroprotective properties of miR-298 and its potential to act at multiple targets (Baza et al.). miR-16 has a similar dual role, in that it reduces APP levels and increases tau levels. miR-16 levels change from early to late-stage AD. This finding indicates that miR-16 may protect in the early stages of the disease, but its expression may be dysregulated as the

pathology progresses. Conversely, miR-125b appears to foster the disease by promoting Aβ accumulation and tau hyperphosphorylation via inflammatory pathways and by regulating kinases. In conclusion, these highly effective dual-acting miRNAs could provide multiple therapeutic options for AD by enabling simultaneous treatment of multiple pathological components of the disease, rather than targeting one pathway at a time.

miRNAs represent emerging candidates as potent diagnostic biomarkers for neurodegenerative diseases such as AD; however, their promise lies in their excellent stabilization and reproducibility in blood (plasma) and spinal fluid (cerebrospinal fluid), and therefore represent excellent candidates for non-invasive detection and monitoring of disease. To date, numerous circulating miRNAs (e.g., miR-125b, miR-132-3p, miR-29b, and miR-107) have been shown to exhibit consistent and measurable changes in patients with mild cognitive impairment (MCI) and early-stage AD, demonstrating their potential as early-stage diagnostic/prognostic markers. In addition to individual markers, a panel of 10 miRNA signatures has been proposed to classify patients as AD or MCI, and studies using longitudinal biofluids suggest that the miRNAs in this panel (e.g., miR-26b, miR-34a, and miR-181c) could potentially detect AD up to 20 years before the onset of clinical symptoms, thereby providing insight into how biomolecular analysis of the brain and spinal fluid may assist physicians in diagnosing AD at an earlier stage. As demonstrated by the accumulation of biomolecular evidence, the proposed miRNA diagnostic test has the potential to enable earlier intervention in the management of AD, thereby promoting better health outcomes and addressing issues related to late diagnosis, which impede the effectiveness of current therapies.

5. Therapeutic Potential of miRNAs

Researchers are investigating whether miRNAs could serve as a new treatment option for patients with AD because they can target multiple genes simultaneously. Strategies for using miRNAs in treatment will include creating synthetic miRNAs that mimic the actions of beneficial miRNAs at appropriate concentrations in the body. Two examples of beneficial miRNAs that reduce Aβ production and decrease tau pathology in vitro are miR-29 and miR-132 (Hebert et al.; Lau et al.). Another method of treating patients with AD is to use antagomirs, or miRNA inhibitors, to block miRNAs involved in the disease's development and progression. For example, this would include inhibiting the action of miR-125b and other similar miRNAs.

6. Conclusion

miRNAs serve a critical role in controlling when and how quickly AD starts and develops, and how well the disease can be treated, through their ability to modulate gene expression in the human brain. Therefore, the interaction of miRNAs with both the Aβ and tau pathways results in alterations in several physiological functions (e.g., protein synthesis, protein clearance, pro-inflammatory mediators, neuronal survival), all of which are represented in the pathology associated with AD. The ability of miRNAs to target multiple genes simultaneously makes them highly regarded, as AD results from many molecular events rather than a single

defect. The discovery of miRNAs in circulating body fluids (e.g., blood, cerebrospinal fluid) has provided evidence of their potential as reliable, non-invasive biomarkers that could enable earlier detection of AD and monitoring of disease progression in ways not yet achieved by existing methods.

Future research must validate miRNA signatures through clinical exploration in larger, more diverse human populations and develop methods for brain-directed delivery systems to improve target-site delivery by overcoming the many obstacles posed by the blood-brain barrier and limiting off-target effects. In addition, a major limitation in this area is that the bulk of evidence to date is from in vitro and animal studies; therefore, well-defined standardized methodologies are still not well established for determining the presence of miRNA within biological specimens and performing comparative analyses across disease states, thereby limiting the ability to translate basic science findings directly into clinical practice.

Works Cited

Banzhaf-Strathmann, Julia, et al. "MicroRNA-125b Induces Tau Hyperphosphorylation and Cognitive Deficits in Alzheimer's Disease." *The EMBO Journal*, vol. 33, no. 15, Aug. 2014, pp. 1667–80. *PubMed*, https://doi.org/10.15252/embj.201387576.

———. "MicroRNA-125b Induces Tau Hyperphosphorylation and Cognitive Deficits in Alzheimer's Disease." *The EMBO Journal*, vol. 33, no. 15, Aug. 2014, pp. 1667–80. *PubMed*, https://doi.org/10.15252/embj.201387576.

Baza, María Begoña, et al. "Natural Conception Is Safe for HIV-Serodiscordant Couples with Persistent Suppressive Antiretroviral Therapy for the Infected Partner." *Journal of Women's Health* [2002], vol. 28, no. 11, Nov. 2019, pp. 1555–62. *PubMed*, https://doi.org/10.1089/jwh.2018.7485.

———. "Natural Conception Is Safe for HIV-Serodiscordant Couples with Persistent Suppressive Antiretroviral Therapy for the Infected Partner." *Journal of Women's Health* [2002], vol. 28, no. 11, Nov. 2019, pp. 1555–62. *PubMed*, https://doi.org/10.1089/jwh.2018.7485.

He, Lin, and Gregory J. Hannon. "MicroRNAs: Small RNAs with a Big Role in Gene Regulation." *Nature Reviews. Genetics*, vol. 5, no. 7, July 2004, pp. 522–31. *PubMed*, https://doi.org/10.1038/nrg1379.

Kavanagh, J. J., et al. "Phase I Study of Fludarabine (2-Fluoro-Ara-AMP)." *European Journal of Cancer & Clinical Oncology*, vol. 21, no. 9, Sept. 1985, pp. 1009–11. *PubMed*, https://doi.org/10.1016/0277-5379(85)90283-4.

Li, Shijie, et al. "The Role of microRNAs in Neurodegenerative Diseases: A Review." *Cell Biology and Toxicology*, vol. 39, no. 1, Feb. 2023, pp. 53–83. *PubMed*, https://doi.org/10.1007/s10565-022-09761-x.

Vernet, Erik, et al. "Screening of Genetic Parameters for Soluble Protein Expression in Escherichia Coli." *Protein Expression and Purification*, vol. 77, no. 1, May 2011, pp. 104–11. *PubMed*, https://doi.org/10.1016/j.pep.2010.11.016.

Woods, R. Jeremy, et al. "LC-MS Characterization and Purity Assessment of a Prototype Bispecific Antibody." *mAbs*, vol. 5, no. 5, 2013, pp. 711–22. *PubMed*, https://doi.org/10.4161/mabs.25488.

Chesnokova, Vera, et al. "Growth Hormone Is Permissive for Neoplastic Colon Growth." Proceedings of the National Academy of Sciences of the United States of America, vol. 113, no. 23, June 2016, pp. E3250-3259. PubMed, https://doi.org/10.1073/pnas.1600561113.

———. "Growth Hormone Is Permissive for Neoplastic Colon Growth." Proceedings of the National Academy of Sciences of the United States of America, vol. 113, no. 23, June 2016, pp. E3250-3259. PubMed, https://doi.org/10.1073/pnas.1600561113.

Hardy, John, and Dennis J. Selkoe. "The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics." Science, vol. 297, no. 5580, July 2002, pp. 353–56. DOI.org (Crossref), https://doi.org/10.1126/science.1072994.

He, Lin, and Gregory J. Hannon. “MicroRNAs: Small RNAs with a Big Role in Gene Regulation.” Nature Reviews Genetics, vol. 5, no. 7, July 2004, pp. 522–31. DOI.org (Crossref), https://doi.org/10.1038/nrg1379.

———. “MicroRNAs: Small RNAs with a Big Role in Gene Regulation.” Nature Reviews. Genetics, vol. 5, no. 7, July 2004, pp. 522–31. PubMed, https://doi.org/10.1038/nrg1379.

Hébert, Sébastien S., et al. “Loss of microRNA Cluster miR-29a/b-1 in Sporadic Alzheimer’s Disease Correlates with Increased BACE1/Beta-Secretase Expression.” Proceedings of the National Academy of Sciences of the United States of America, vol. 105, no. 17, Apr. 2008, pp. 6415–20. PubMed, https://doi.org/10.1073/pnas.0710263105.

Hébert, Sébastien S., and Bart De Strooper. “Alterations of the microRNA Network Cause Neurodegenerative Disease.” Trends in Neurosciences, vol. 32, no. 4, Apr. 2009, pp. 199–206. DOI.org (Crossref), https://doi.org/10.1016/j.tins.2008.12.003.

Johnson, A. D., et al. “Genome-Wide Association Meta-Analysis for Total Serum Bilirubin Levels.” Human Molecular Genetics, vol. 18, no. 14, July 2009, pp. 2700–10. DOI.org (Crossref), https://doi.org/10.1093/hmg/ddp202.

Kavanagh, J. J., et al. “Phase I Study of Fludarabine (2-Fluoro-Ara-AMP).” European Journal of Cancer & Clinical Oncology, vol. 21, no. 9, Sept. 1985, pp. 1009–11. PubMed, https://doi.org/10.1016/0277-5379(85)90283-4.

Kling, Mitchel A., et al. “Vascular Disease and Dementias: Paradigm Shifts to Drive Research in New Directions.” Alzheimer’s & Dementia, vol. 9, no. 1, Jan. 2013, pp. 76–92. DOI.org (Crossref), https://doi.org/10.1016/j.jalz.2012.02.007.

Kosik, Kenneth S. “The Neuronal microRNA System.” Nature Reviews Neuroscience, vol. 7, no. 12, Dec. 2006, pp. 911–20. DOI.org (Crossref), https://doi.org/10.1038/nrn2037.

Pajoni, D. “[Tomodensitometry and dentistry].” Revue D’odonto-Stomatologie, vol. 20, no. 4, 1991, pp. 265–78.

Parvand, Mahraz, and Catharine H. Rankin. “Is There a Shared Etiology of Olfactory Impairments in Normal Aging and Neurodegenerative Disease?” Journal of Alzheimer’s Disease, edited by Lori Beason-Held, vol. 73, no. 1, Jan. 2020, pp. 1–21. DOI.org (Crossref), https://doi.org/10.3233/JAD-190636.

Querfurth, Henry W., and Frank M. LaFerla. “Alzheimer’s Disease.” New England Journal of Medicine, vol. 362, no. 4, Jan. 2010, pp. 329–44. DOI.org (Crossref), https://doi.org/10.1056/NEJMra0909142.

———. “Alzheimer’s Disease.” The New England Journal of Medicine, vol. 362, Jan. 2010, pp. 329–44. PubMed, https://doi.org/10.1056/NEJMra0909142.

Sale, Alessandro, et al. “Enrich the Environment to Empower the Brain.” Trends in Neurosciences, vol. 32, no. 4, Apr. 2009, pp. 233–39. PubMed, https://doi.org/10.1016/j.tins.2008.12.004.

Chemical Strategies for Mitigating Jeju Island's Groundwater Contamination
By Hyunsung Ko

Introduction

Jeju Island, a UNESCO World Natural Heritage site located off the southern coast of South Korea, is renowned for its pristine beaches, volcanic landscapes, and unique biodiversity. As one of South Korea's most popular tourist destinations and an important ecological region, Jeju's coastal and marine environments play a pivotal role in both the local economy and the conservation of globally significant species.

Access to clean and sustainable water resources is essential not only for the public health, but also for maintaining ecological stability, agricultural productivity, and economic development in island environments such as Jeju. Jeju Island already operates a network of water purification and wastewater treatment facilities managed by the Jeju Special Self-Governing Province Waterworks and Sewerage Headquarters. Currently, there are 17 water purification plants with several public sewage treatment plants. These plants treat tens of thousands of tons of sewage per day. Alongside these plants, Jeju also heavily relies on individual sewage treatment facilities and septic systems, with over 10,000 small on-site systems operating across the island (Park, 2025).

However, Jeju has faced countless challenges related to water and ocean pollution (Choi, 2006). Jeju Island relies on groundwater for more than 90% of its total water resources, which accounts for more than 96% of water for agriculture (JSSGP, 2018). Rapid urbanization, expanded tourism infrastructure, increased maritime traffic, and intensified agricultural activities have also contributed to the pollution of the island's freshwater and marine systems. These factors have led to escalating waste, such as plastics and microplastics, which disrupt marine habitats.

Jeju Island's groundwater is increasingly contaminated by both agricultural runoff and microplastics due to the fact that more than 90 percent of its water supply comes from highly permeable basalt aquifers that are exposed to human activity (Kim, 2023). Agricultural practices such as greenhouses and fertilizer use have been identified as major sources of microplastic contamination. Plastic greenhouse covers, irrigation tubing, and mulch films gradually degrade through sunlight, exposure, and weathering, releasing microplastic particles into surrounding soils and groundwater systems. Recent work has also confirmed that microplastic occurs throughout Jeju's groundwater system, dominated by common consumer polymers such as polypropylene and polyethylene. These are mainly from domestic wastewater, laundry effluent, and tourism-related wastes. Together, these findings suggest that Jeju's contamination is already threatening the drinking water supply, agriculture, and coastal ecosystems.

Similar to Jeju Island, Dawanshan Island in the South China Sea is also facing serious microplastic pollution. Its similar reliance on groundwater, limited industrial activity, and vulnerability to domestic pollution sources make Dawanshan Island a compelling comparative case study for Jeju Island. A 2023 study found that microplastic activity in the Dawanshan

island's groundwater ranged from 34 to 64 particles per liter, indicating significant subsurface pollution in areas with minimal agriculture or heavy industry (Gong et al., 2023). Globally, compiled groundwater datasets show microplastic abundances around 2.5-4.4 items per year (He et al., 2024). These comparative numbers demonstrate how heavily polluted the island is. The major microplastics causing the pollution were polyester fibers and fragments, originating from domestic wastewater, laundry effluents, and local plastic wastes.

This case parallels Jeju's challenges. The majority of the water resources come from volcanic groundwater, which is susceptible to tourism-related waste and agricultural system byproducts. For example, Jeju's microplastic levels of 0.006-0.192 particles per liter are much lower than both the global groundwater median and Dawanshan's extreme environment (Kim, 2023). Lower concentrations may sound better, but Jeju's levels being below the global median could indicate that it's at early-stage contamination and still raises safety concerns.

Recognizing the urgency of this issue, significant research and policy efforts are being made to identify pollution sources and find strategies to mitigate the impact. This literature review addresses these gaps by pursuing three specific objectives: (1) characterize Jeju Island's groundwater contamination profile; (2) evaluate three chemical treatment technologies; and (3) propose a prioritized implementation strategy. Among these strategies, chemical water treatment methods play a pivotal role in removing harmful contaminants before they reach sensitive ecosystems. While these methods are widely used for water treatment around the world, understanding the effectiveness and limitations for Jeju's specific context is essential.

Methodology

The review draws from peer-reviewed journal articles, government and environmental agency reports, and case studies focusing on both Jeju-specific contexts and comparable general ecosystem information. Focusing on studies on microplastic contamination in volcanic groundwater, the primary step involves searching academic databases such as Google Scholar and JSTOR, using targeted keywords like "Jeju Island water pollution", "Jeju groundwater contamination", and "ocean pollution Jeju tourism". Selected research was published within the last 10-15 years, providing reliable sources for solutions.

Available or Commonly-Used Water Treatment Technologies in Developed Nations

Chemical water treatment methods comprise an essential approach to managing water pollution by removing or neutralizing contaminants. The three major categories of chemical treatments include coagulation and precipitation, adsorption, and advanced oxidation processes (AOPs) (Kim, 2023). Coagulation and precipitation involve the addition of chemical coagulants such as alum and ferric salts to water, which destabilize suspended particles and dissolved substances. These particles aggregate to form precipitates that can be physically removed. This technique is widely employed in municipal and industrial water treatment facilities worldwide, including plants in Korea, South Africa, and the United States. Coagulation effectively reduces turbidity, color, and certain dissolved pollutants, including heavy metals. However, this method

generates chemical sludge that requires proper disposal and can alter the water's pH, necessitating careful dosing to prevent residual chemical contamination.

Adsorption utilizes porous materials like activated carbon and iron oxide to attract and bind pollutants on their surface, thereby removing organic micropollutants, heavy metals, and other inorganic substances. This method is commonly used for treating drinking water and industrial wastewater globally, with activated carbon filters implemented in municipal systems and iron-based adsorbents used specifically to remove arsenic in parts of Asia and the United States. The adsorption technique is highly efficient for organic compounds and relatively simple to operate. However, the materials used as adsorbents require regeneration or replacement, as their effectiveness declines when saturated. Additionally, the presence of competing substances in the water can reduce adsorption efficiency, making pre-treatment steps, such as coagulation or filtration, beneficial in maintaining performance (Rajamohan, 2025; Verma, 2021).

AOPs employ reactive species like hydroxyl radicals generated through ozone, hydrogen peroxide, UV light, or in combination. These highly reactive radicals break down complex organic pollutants that are difficult to treat with conventional methods. AOPs are increasingly applied in developed countries for the removal of refractory organics, pharmaceuticals, and various industrial contaminants. The primary advantage of AOPs is their ability to degrade a wide range of persistent pollutants and effectively kill pathogens. However, high energy consumption, operational costs, the need for specialized equipment, and potential formation of harmful byproducts present significant limitations for widespread adoption (Ultraqua 2025).

When comparing these three methods, each has unique strengths and weaknesses. Coagulation and precipitation are effective for removing particulate matter and some dissolved substances, but the production of sludge raises environmental and disposal concerns. It contains concentrated coagulant chemicals, aggregated particles, and trapped contaminants that must be treated and safely disposed of to prevent secondary environmental pollution. Adsorption is well-suited for targeting a broad range of organic pollutants with simpler operational requirements, although the necessity for regular adsorbent replacement adds ongoing costs. AOPs offer superior treatment for persistent and emerging contaminants, but at the expense of greater complexity, higher costs, and increased energy demands. From a resource perspective, coagulation and adsorption require moderate space and training, while AOPs necessitate more specialized knowledge, infrastructure, and safety protocols. Environmentally, each method impacts water treatment systems differently: coagulation generates sludge and alters chemical balances; adsorption produces spent adsorbents requiring disposal; and AOPs risk byproduct formation and higher carbon footprints due to energy use (Perumal, 2024).

In conclusion, chemical water treatment methods play a vital role in addressing the diverse causes of water pollution worldwide. Coagulation and precipitation serve as reliable preliminary treatments for suspended and dissolved contaminants, but must be managed to mitigate sludge impacts. Coagulation is already being used in many drinking-water treatment plants worldwide, with coagulants such as alum used in roughly 44% of facilities (360 Research Reports, 2024).

Potential Solutions and Feasible Implementations for Jeju Island

Jeju Island's water resources are vulnerable to contamination from agriculture and tourism, and limited groundwater recharge due to porous volcanic geology. Adsorption via activated carbon is feasible, especially as a tertiary treatment, which is an advanced final purification step used after conventional wastewater treatment to remove remaining contaminants, or for decentralized systems where groundwater is contaminated with organics or metals. The cost is moderate—particularly if low-cost biosorbents (from local agricultural waste) are tested. Maintenance involves periodic replacement or regeneration of the adsorbent, which is manageable and supported by Korean environmental standards.

Similar to adsorption, coagulation can be directly applied to the island. It could provide efficient removal of pollutants and particulates such as agricultural runoff and microplastics in Jeju groundwater. Coagulation and precipitation methods remain the most feasible for immediate improvements, given their lower operational costs, maturity, and proven success in small system contexts like Jeju. The equipment and chemicals required are cost-effective and do not require specialized skills, making rollout and maintenance relatively easy. These techniques are already being used in many water systems across Korea and on Jeju to remove particulate phosphates and other matter. The main downside is the disposal of sludge byproducts, which contain concentrated coagulant chemicals and trapped contaminants and must be treated and disposed of carefully to avoid environmental pollution (Alkhadra et al., 2022).

AOPs also show potential for addressing persistent and trace organic contaminants, which are increasingly detected in other island water systems with tourism and development. AOPs utilize physical or chemical technologies like ozone or UV/hydrogen peroxide, which offer the most comprehensive treatment; their costs, electricity demands, and required infrastructure make them less immediately feasible for widespread rollout. For targeted "hot spots" near resorts or industrial discharge points, however, pilot projects could demonstrate their benefits, particularly as Jeju looks to future-proof its water supply and protect against emerging contaminants (Hubner et al., 2024). However, higher energy is required and the need for technical expertise might limit the adoption of AOP unless external funding is available.

Implementing any of the technologies directly tackles Jeju's core issues. Cost-effective rollout starts with upgrading existing Jeju plants, in other words, coagulation enhancements. Those power plants include the Thermal Power Station, which has fuel-oil units and internal combustion power plants. Then, piloting absorption methods using inexpensive local materials, such as fruit waste, would save nearly half of the money on purchasing materials from outside.

Conclusion

Jeju Island's water and ocean pollution crisis, driven by groundwater microplastic contamination, tourism sewage, and agricultural runoff, demands a chemical solution to protect its UNESCO heritage. Coagulation/precipitation and adsorption emerge as the most promising and feasible solutions for their low costs, rapid deployment, and proven applicability. AOPs

represent feasible solutions, but they are constrained by unique challenges that require additional time and financial investment.

Ultimately, monitoring, local research, and partnerships among farmers, tourism operators, and policymakers, are vital to refine these strategies against environmental threats. By prioritizing feasible chemical solutions informed by global and Jeju-specific evidence, the island can restore its original water quality, upholding its status as a global natural treasure with the title “UNESCO Global Geopark”.

Works Cited

Choi, Yearn Hong, and Choong Suk Ko. "JST." Environmental Issues in Jeju Island, Korea., jst.re.kr/upload/pdf/RC00006787.pdf.

Cui, Hongmei et al. "Application progress of enhanced coagulation in water treatment." RSC advances vol. 10,34 20231-20244. 27 May. 2020, doi:10.1039/d0ra02979c

Hajalifard, Z., Mousazadeh, M., Khademi, S. et al. The efficiency of AOP-based processes in concert with electrocoagulation in abatement of CECs from water/wastewater. npj Clean Water **6**, 30 (2023). https://doi.org/10.1038/s41545-023-00239-9

Hajalifard, Zeinab, et al. "The Efficacy of AOP-Based Processes in Concert with Electrocoagulation in Abatement of Cecs from Water/Wastewater." Nature News, Nature Publishing Group, 11 Apr. 2023, www.nature.com/articles/s41545-023-00239-9.

Hübner, Uwe et al. "Advanced oxidation processes for water and wastewater treatment - Guidance for systematic future research." Heliyon vol. 10,9 e30402. 28 Apr. 2024, doi:10.1016/j.heliyon.2024.e30402

Inorganic Coagulants Market Size, Share, Growth, and Industry Analysis, by Type (Aluminum Sulfate, Polyaluminum Chloride, Ferric Chloride, Ferrous Sulfate, Others), by Application (Municipal Water Treatment, Sewage Systems, Industrial Water Treatment, Paper Industry, Oil&gas, Others), Regional Insights and Forecast to 2035." Market Reports, www.marketgrowthreports.com/market-reports/inorganic-coagulants-market-117373. Accessed 8 Mar. 2026.

Jensby, Andreas. "UV & Ozone Treatment for Sustainability - World Water Day 2025." ULTRAAQUA, 14 Nov. 2025, ultraaqua.com/blog/world-water-day-2025/.

Jinkeun Kim. Current status of Jeju special self-governing province's water infrastructure and direction for improvement. Journal of the Korean Society of Water and Wastewater, 35.6 (2021):497-505.

Kim, Young-In, et al. "Microplastic Contamination in Groundwater on a Volcanic Jeju Island of Korea." Environmental Research, vol. 226, 2023, article 115682

Satyam, Satyam, and Sanjukta Patra. "Innovations and challenges in adsorption-based wastewater remediation: A comprehensive review." Heliyon vol 109, 10,9 e29573. 20 Apr. 2024, doi:10.1016/j.heliyon.2024.e29573

Park, Changwon. "Jeju Waterworks and Sewerage Headquarters Selected as 'Best Water Management Institution.'" The Asia Business Daily, 13 Feb. 2025, www.asiae.co.kr/en/article/2025021317141524191.

Housing Stability and Access to Opportunity: A Comparative Study of Low-Income Housing Policy and Urban Planning in New York City, Hong Kong, and Tokyo

By Zhixiao Max Liu

Abstract

New York City, Hong Kong, and Tokyo are dense global cities facing high housing costs, limited land, and strong demand for urban space, but they differ substantially in how housing policy, land-use planning, infrastructure, and housing supply interact. This paper compares the three cities to examine how different combinations of housing assistance, public-housing availability, and transit structure are associated with low-income households' housing stability and access to employment.

The comparison shows that low-income households face the greatest instability when they remain exposed to the private rental market, but the cause and extent of that exposure differ across cities. In Hong Kong, the main challenge is waiting time: public rental housing provides deep affordability and generally moderate commute times after placement, but many households must first spend years in the private market before receiving that protection. In Tokyo, the main challenge is scale. The city's supply-oriented housing system and rail-based urban structure support broader housing availability and relatively strong employment access, but its deeply affordable public housing sector remains too small to serve most lower-income households. In New York City, both problems appear together. Multiple housing programs, including public housing, rental subsidies, and rent regulation, reduce instability for some households, but limited availability, long waitlists, shrinking regulated stock, and long commutes from many lower-rent areas leave many households exposed. Overall, the paper suggests that low-income housing policy should be evaluated not only by rent levels or unit counts, but by whether affordable housing is available when households need it and whether it connects them to employment opportunities.

1 Introduction

Housing affordability is not only a question of rent levels; for low-income households, it is closely tied to whether affordable units are available when needed and whether those units are connected to jobs and other opportunities. A household may qualify for public or subsidized housing but still face years of waiting; another may find cheaper housing only by accepting a much longer commute. These trade-offs show why low-income housing policy should be studied alongside urban planning and transit access.

This paper compares New York City, Hong Kong, and Tokyo to examine how different combinations of housing assistance, public-housing availability, and transit structure are connected to low-income households' housing stability and access to employment.

2 Data and Method

Using government reports, official statistical publications, housing surveys, and author-calculated transit and rent indicators, this study evaluates two outcomes: housing stability and access to opportunity. Table 1 summarizes the comparative measures used in the study. Housing stability is assessed through housing affordability, rent burden, and access delay. Access to opportunity is assessed through public-transit access and the location of affordable housing relative to major employment centers.

Table 1. Comparative Measures Used in the Study

Outcome	Concept	NYC measure	Hong Kong measure	Tokyo measure
Housing stability	Housing affordability and rent burden	NYCHVS rent-to-income by income group; median rent by housing type	Public/private rent per sqm; estimated public vs private rent burden	Average rent by tenure; estimated rent burden by tenure and income band
Housing stability	Access delay	NYCHA waitlists, turnover, Section 8 status	Average public rental housing waiting time; applicants housed	Toei Housing application ratios
Access to opportunity	Transit access	Author-calculated public-transit travel bands to job centers	Mean public-transport journey time	Station distance and commute time by housing type
Access to opportunity	Affordable-housing location	Rent by travel-time band; under-$600 rent tract map	Public housing distribution by region	Median commute time by tenure

Because the three cities report housing data differently, the paper compares functionally equivalent indicators rather than forcing identical measures across all cases. Rent burden is defined as the share of household income spent on rent. Access delay is measured using the available indicators for each city, including waitlist length, average waiting time, housing turnover, and application ratios. Public-transit access is defined as connection to major employment centers, measured through travel time, journey time, station proximity, or commute time depending on available data. Affordable-housing location is assessed by examining whether lower-cost or public housing is concentrated closer to or farther from major employment centers. When indicators differ across cities, the paper notes the interpretive limits of the comparison.

3 New York City

3.1 Housing Stability

New York City's housing support system relies on multiple programs that address different housing needs simultaneously, rather than centering on a single dominant form of assistance. Within this system, public housing operated by the New York City Housing Authority (NYCHA) offers the deepest long-term stability because it provides permanently affordable units. The Section 8 Housing Choice Voucher program and the city's CityFHEPS rental subsidies help households remain in privately owned apartments by reducing rent burdens and lowering

the immediate risk of housing loss.[1] Rent stabilization and rent control both protect existing tenants by limiting annual rent increases and strengthening renewal rights.[2] Housing Connect provides access to newly created affordable units through competitive lotteries. Each program supports housing stability in a different way, but access to these programs is uneven: some households receive durable public housing, while many others depend on waiting lists, landlord participation, or lottery selection.

Longer-term changes in the rental stock, shown in table 2, reveal why access to stable housing has become more constrained. From 2011 to 2023, total regulated rentals declined from 1,356,111 to 1,252,130, reducing the relative size of the more protected segment of the rental market. Although market rental units increased from 816,522 to 1,139,000 over the same period, this growth occurred outside the more protected housing categories. As a result, the overall rental stock expanded, but the segment most likely to provide long-term affordability and stability became relatively more limited.

Table 2. Rental Units by Type of Housing

	2011	2014	2017	2021	2023
Market Rentals	816,522	848,721	936,850	1,023,000	1,139,000
Total Regulated Rentals	1,356,111	1,335,657	1,246,220	1,251,260	1,252,130
Rent Stabilized	*1,020,727*	*1,030,000*	*966,442*	*1,006,000*	*996,600*
Public Housing + Other Regulated Rental	*335,384*	*305,657*	*279,778*	*245,260*	*255,530*

Note: Category definitions are not fully consistent across survey years. In the 2017 NYCHVS, "all other rental units" included public housing, Mitchell-Lama, in rem, HUD-regulated, Article 4, Municipal Loan, and Loft Board units. In the 2021 and 2023 NYCHVS, public housing is reported separately, while "other rentals" excludes public housing. Total regulated rentals are calculated by adding rent-stabilized units and public housing/other regulated rental units.
Source: Author's table using data from New York City Department of Housing Preservation and Development, New York City Housing and Vacancy Survey: Selected Initial Findings, Table 3 for 2011, 2014, and 2017; Figure 2 for 2021 and 2023.

Table 3 shows why the decline in regulated rental stock matters for low-income households. Different forms of regulated housing provide very different degrees of affordability. Between 2021 and 2023, the citywide median rent increased from $1,500 to $1,641, or about 9 percent. Public housing remained much more affordable, with a median rent of $560 in 2023, about one-third of the citywide median. Rent-controlled units also remained below the citywide level, at $988, while rent-stabilized units had a median rent of $1,500, only slightly below the citywide median. These differences show that regulated housing can support stability, but the depth of affordability varies sharply across programs.

[1] Section 8 refers to the federal Housing Choice Voucher program, which subsidizes rent in privately owned housing. CityFHEPS is New York City's local rental assistance program for eligible households at risk of or experiencing homelessness.
[2] Rent stabilization is the larger and more common form of rent regulation in New York City, applying to many older rental buildings and some newer subsidized or tax-benefit buildings. Rent control is an older and much smaller category, generally applying to long-term tenants in older apartments. Both limit rent increases and provide protections against displacement, but rent-controlled units are far less common than rent-stabilized units.

Table 3. Median Rent by Type of Housing

	2011	2014	2017	2021	2023
Citywide	$1,100	$1,200	$1,337	$1,500	$1,641
Market Rentals	$1,369	$1,500	$1,700	$1,825	$2,000
Rent Stabilized	$1,050	$1,200	$1,269	$1,400	$1,500
Public Housing	$591	$583	$610	$500	$560
Rent Controlled	$800	$900	$915	$858	$988

Source: Author's table using data from New York City Department of Housing Preservation and Development, New York City Housing and Vacancy Survey: Selected Initial Findings, Table 12 for 2011, 2014, and 2017; Table 6 for 2021 and 2023.

Even where regulated housing is more affordable, availability remains constrained by very low vacancy rates. As Figure 1 shows, by 2023, vacancy rates had fallen to just 0.98% for rent-stabilized units and 1.21% for public housing and other regulated units, indicating that the more affordable and protected segments of the rental market had very few available homes.

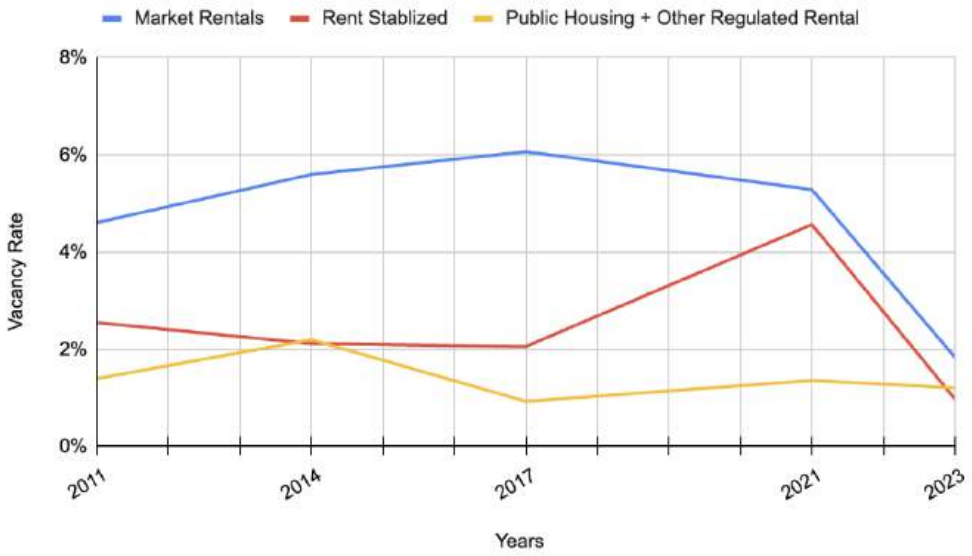


Figure 1. Vacancy Rates by Type of Housing

Source: Author's chart using data from New York City Department of Housing Preservation and Development, New York City Housing and Vacancy Survey: Selected Initial Findings, Table 6 for 2011, 2014, and 2017; Table 9 for 2021 and 2023.

Low vacancy in regulated housing indicates that affordable units were difficult to obtain. Scarcity was also visible in the private rental market. As Figure 2 shows, only 4,442 vacant units available for rent in 2023 were affordable to households earning around $50,000, compared with substantially larger numbers affordable only at higher income levels. This suggests that the private rental market offered very few available options for lower-income households.

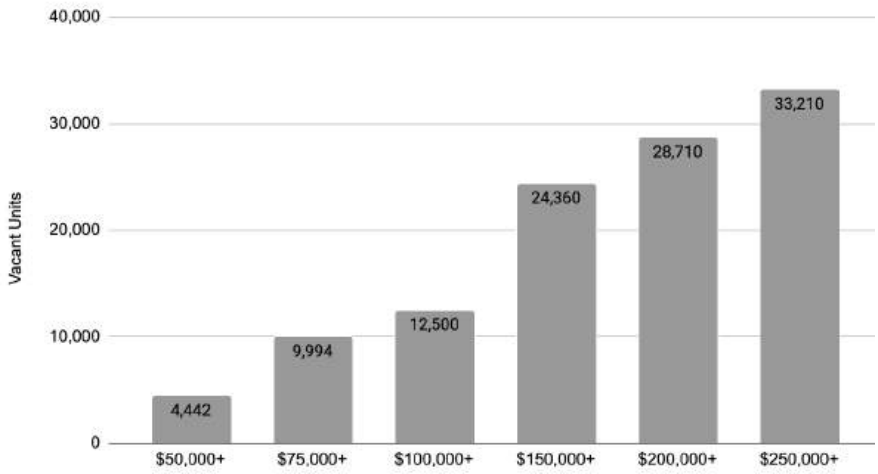


Figure 2. Cumulative Number of Vacant Units Affordable to Households at Different Income Thresholds

Source: Author's chart using data from New York City Department of Housing Preservation and Development, *New York City Housing and Vacancy Survey: Selected Initial Findings*, 2023, Figure 9.

Figure 3 shows that the rent burden was especially severe for low-income renters who remained in the private rental market. The 2023 New York City Housing and Vacancy Survey (NYCHVS) classified households into three levels of rent burden: severely rent burdened, meaning they paid more than 50 percent of household income toward rent; moderately rent burdened, meaning they paid between 30 and 50 percent; and not rent burdened, meaning they paid 30 percent or less. According to the survey, renters earning less than $70,000 had a median rent-to-income ratio of 54 percent in 2023, above the threshold for severe rent burden.

The pressure was greatest for households earning below $50,000. Among households earning less than $25,000 and not living in public housing or reporting a voucher, 86 percent were severely rent burdened. Among households earning between $25,000 and $49,999, 45 percent were severely rent burdened.

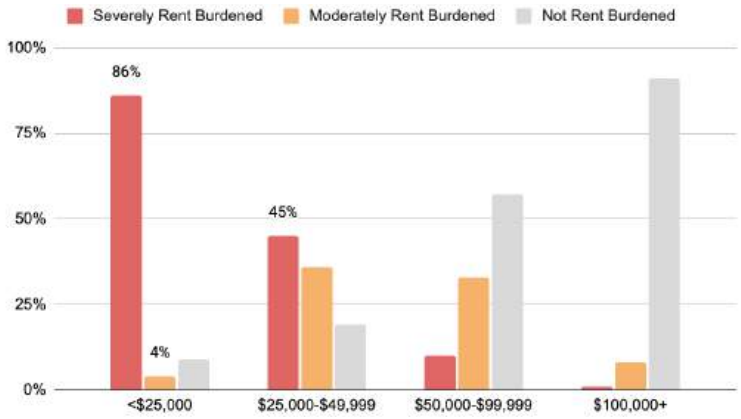


Figure 3. Rent Burden, by Household Income, New York City, 2023

Source: Author's chart using data from New York City Department of Housing Preservation and Development, *New York City Housing and Vacancy Survey: Selected Initial Findings*, 2023, Figure 12.

These affordability pressures help explain why demand for subsidized assistance far exceeds available support. As shown in Table 4, as of March 3, 2025, the public housing waiting list included 179,664 families, while the Section 8 tenant-based assistance waiting list included 198,657 families. NYCHA public housing also had low turnover, with the most recent published annual turnover rate at 2.2 percent in 2024, limiting the speed at which applicants could be housed. Although NYCHA briefly reopened the Section 8 list on June 3, 2024, and replenished it with 200,000 applicants, the list was closed again to the general public on June 9, 2024. This brief reopening shows that access to voucher assistance remains intermittent and highly constrained.

Table 4. Waiting Lists for Deeply Affordable Housing Assistance, New York City, 2025

Program	Families on waiting list	Status / constraint
NYCHA Public Housing	179,664	Low turnover limits movement from the waitlist
NYCHA Section 8	198,657	General waitlist closed to new applicants

Source: Author's table using data from New York City Housing Authority, *Draft Public Housing Agency Plan: Draft Annual Agency Plan for Fiscal Year 2026*, public housing waiting-list and turnover-rate data as of March 3, 2025, p. 185; Section 8 waiting-list data, pp. 186–187.

At the same time, NYCHA faces a severe capital constraint. Between 2001 and 2023, NYCHA experienced a cumulative loss of $232 million in federal capital grant funding.[3] In its 2023 Capital Plan for 2024–2028, NYCHA estimated $78.3 billion in 20-year physical needs across the 264 developments it directly manages, totaling roughly 161,400 apartments. This scale of unmet need leaves NYCHA without sufficient capital to fully preserve its existing housing stock, let alone build new public housing at a scale large enough to reduce waiting-list pressure. In response, the city has relied more heavily on affordable-housing finance, preservation, and inclusionary housing programs to maintain or add affordable units. In FY2024, the New York City Department of Housing Preservation and Development (HPD) and the New York City Housing Development Corporation (HDC) financed 25,266 affordable homes, including 14,706 newly constructed affordable homes.[4] The city also produced 5,401 421-a standalone affordable units through a property-tax incentive and 3,255 permanently affordable inclusionary housing units through zoning-related programs.[4] Although these programs expand or preserve affordability, their scale remains limited compared with the city's overall housing need.

In the absence of sufficient new deeply affordable housing, New York City has increasingly relied on rental subsidies and tenant protections to prevent displacement. CityFHEPS, the nation's largest municipally funded rental subsidy program, served 64,439 cases as of October 2025, with an average monthly subsidy of $1,973.[5] Through its 2017 Right to Counsel law, the city also provides legal support to tenants facing eviction. In FY2024, 89

[3] New York City Housing Authority, *2024–2028 Capital Plan Narrative*, 2024, pp. 5, 8.
[4] New York City Department of Housing Preservation and Development, "New York City Announces Back-to-Back Record Years for Creating and Connecting New Yorkers to Affordable Housing," July 29, 2024.
[5] New York City Mayor's Office of Operations, *Preliminary Mayor's Management Report, Fiscal Year 2026: Human Resources Administration*, 2026, p. 204.

percent of tenants who received full legal representation remained stably housed.[6] These interventions reduce displacement risk, but they mainly manage housing instability rather than resolve the underlying shortage of deeply affordable units.

New York City's affordable and regulated rental stock is limited in availability and uneven in its degree of affordability and protection. The programs offering the deepest assistance face long waiting lists, low turnover, and major capacity limitations. The city's multi-program system reduces housing insecurity for some low-income households, but it does not provide a reliable path to stable, deeply affordable housing for those who need it most.

3.2 Access to Opportunities

New York City's subway network is organized like a tree branching outward from Manhattan. The strongest corridors run north-south through Manhattan before spreading outward into Brooklyn, Queens, and the Bronx. Many lines are oriented toward Manhattan, with relatively few strong cross-borough connections. As a result, many outer-borough residents depend on a bus connection or a long walk before reaching the subway, adding time and complexity to commutes that appear direct on a map.

To assess transit access, census tracts were grouped into 15-, 30-, 60-, 90-and 120-minute public-transit travel bands from the nearest of two major employment centers: Midtown Manhattan, and Lower Manhattan. Mean gross rent was estimated by dividing aggregate gross rent from ACS Table B25065 by the aggregate number of renter-occupied housing units from ACS Table B25003 within each travel-time band.

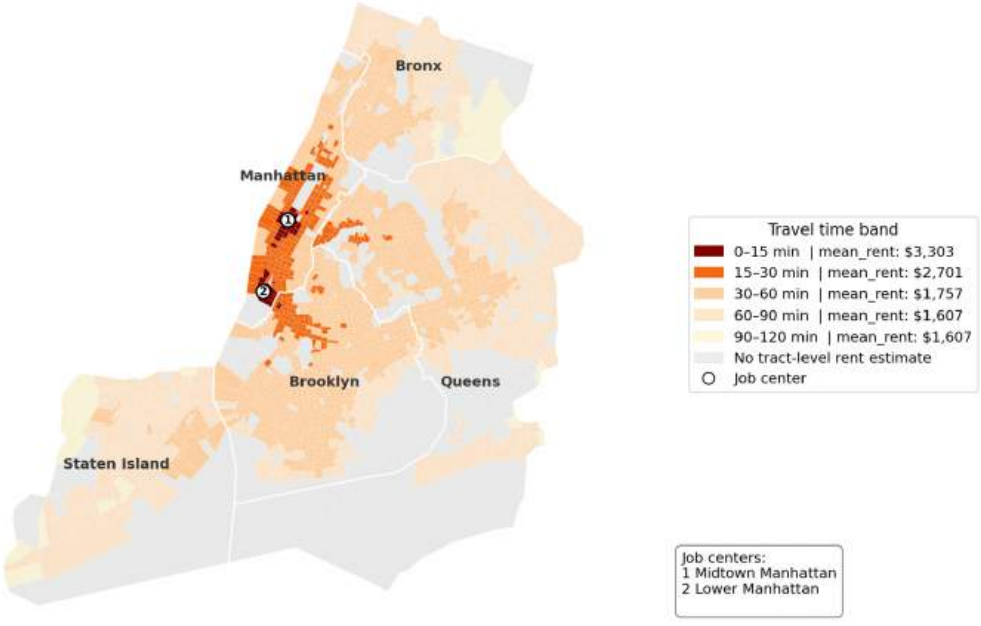


Figure 4. New York City Housing Access to Opportunity: Public-Transit Travel Time to the Nearest Major Job Center

[6] Office of the New York City Comptroller Brad Lander, *Evictions Up, Representation Down: How New York City Is Undermining the Right to Counsel*, May 2025, 4.

Note: Travel-time bands are calculated from each census tract to the nearest of two major employment centers using r5py, a Python interface to the R5 multimodal routing engine. The routing network combines OpenStreetMap street-network data with MTA GTFS feeds for subway, bus, Long Island Rail Road, and Metro-North service. Gray tracts indicate areas without a usable tract-level ACS gross rent estimate in the final rent-access dataset. Figure 4 uses estimated mean gross rent because aggregate gross-rent data allow comparison across travel-time bands.
Source: Author's map using U.S. Census Bureau, American Community Survey 2023 5-Year Estimates, Tables B25065 and B25003; U.S. Census Bureau, 2023 TIGER/Line Shapefiles; Metropolitan Transportation Authority, GTFS subway, bus, Long Island Rail Road, and Metro-North feeds; OpenStreetMap street-network data.

The results, as shown in Figure 4, reveal a clear affordability-access trade-off. Tracts within 15 minutes of a major job center had the highest mean gross rent, at $3,303, compared with $2,701 in the 15–30 minute band. Mean rent was lower in the outer bands, at about $1,757 in the 30–60 minute band and $1,607 in the 60–120 minute band. These patterns reflect both transit structure and housing composition. Many lower-income households live in outer-borough neighborhoods where subway access is indirect, express service is limited, or a bus connection is required before reaching the train. The similarity between the 60–90 and 90–120 minute rent bands should also be interpreted cautiously because ACS median gross rent does not control for unit size, housing type, or housing quality. Households farther from job centers may therefore pay similar rents in exchange for larger units or different housing stock, meaning the trade-off is not only between rent and commute time, but also between commute time and housing size.

To examine the deepest affordable segment more closely, Figure 5 narrows the focus to census tracts with median gross rent below $600, a threshold chosen because it is close to the 2023 median monthly rent in public housing ($560).

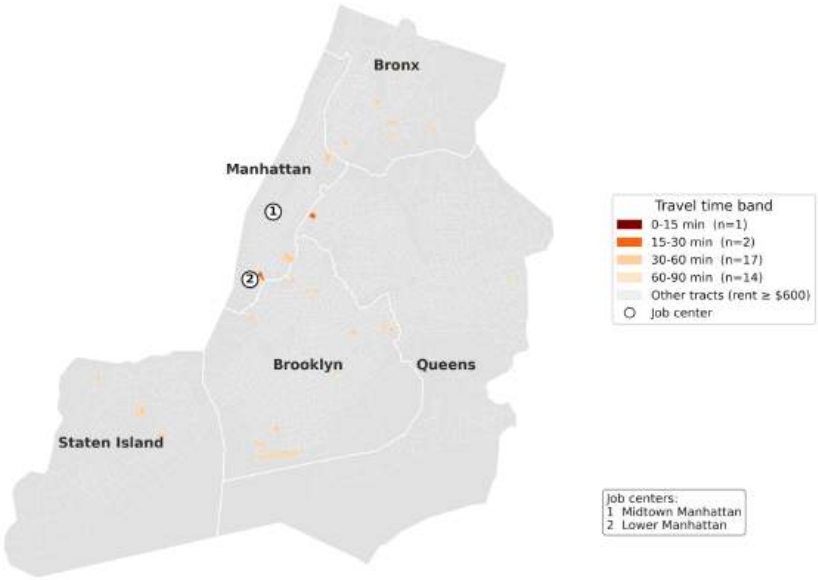


Figure 5. New York City Census Tracts with Median Gross Rent Below $600 by Public-Transit Travel Time to the Nearest Major Job Center

Note: Travel-time bands are calculated using the same r5py-based routing method described in Figure 4. The $600 threshold is used because it is close to the 2023 median monthly rent in public housing ($560). Colored tracts indicate census tracts whose ACS 2023 5-year median gross rent falls below $600; colors represent public-transit travel time to the nearest of two major employment centers. Because ACS median gross rent is a tract-level measure, it does not identify specific housing programs or

control for unit size and housing quality. Figure 5 uses median gross rent because the purpose is to identify tracts whose typical rent level approximates public-housing affordability.

Source: Author's map using U.S. Census Bureau, American Community Survey 2023 5-Year Estimates, Table B25064; U.S. Census Bureau, 2023 TIGER/Line Shapefiles; Metropolitan Transportation Authority (MTA), GTFS subway, bus, Long Island Rail Road, and Metro North feeds.

Most such tracts fall in the 30–60 minute and 60–90 minute public-transit bands from the nearest major job center, while only a very small number are located within 30 minutes. This distribution indicates that the lowest-rent housing is generally located farther from the city's main employment concentrations, suggesting that deeply affordable housing in New York City often comes with weaker access to employment opportunities.

The analysis of housing stability and access to opportunity shows the limits of the city's multi-program assistance system operating in a deeply unaffordable market with uneven transit accessibility. Public housing, rental vouchers, and rent regulation improve housing stability for some households, but low vacancy, shrinking regulated stock, and long delays in obtaining housing assistance mean that support does not reach enough households in time. At the same time, affordability often comes at the cost of longer commute times and weaker access to employment, reinforcing spatial inequality across the urban housing system.

4 Hong Kong

4.1 Housing Stability

Hong Kong provides a comparative case in which low-income housing support centers on public rental housing rather than a multi-program assistance system. The Housing Bureau's *Housing in Figures 2025* reports that, as of March 31, 2025, 45.4% of domestic households lived in public rental housing, compared with 53.8% in private permanent housing and 0.8% in temporary housing. These figures show that public rental housing is not marginal, but a central component of Hong Kong's housing system.

However, the existence of a large public rental housing system does not guarantee timely access. As Figure 6 shows, average waiting times for general applicants remained persistently high, ranging from 4.7 to 6.0 years between 2016 and 2025 and still standing at 5.6 years in 2025. Although the number of general applicants declined over this period, it consistently remained above 100,000. During the same period, the number of general applicants housed each year never reached 30,000, with the highest annual total at about 26,400 in 2022[7]. These figures suggest that demand remained significantly higher than timely supply. As a result, many low-income households remain unable to obtain the stability that public rental housing is meant to provide when they need it most.

[7] Housing Bureau, Hong Kong SAR Government, *Composite Waiting Time for Subsidised Rental Housing*, 2025, Figure on general applicants housed and average waiting time.

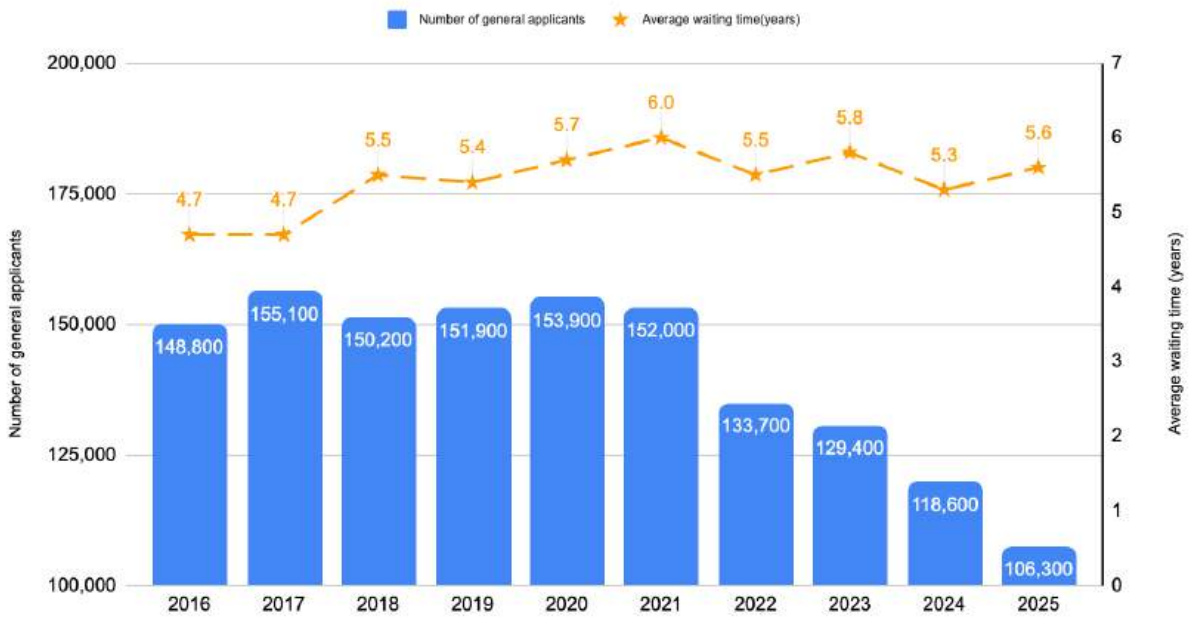


Figure 6. General Applicants and Average Waiting Time for Public Rental Housing, Hong Kong, 2016–2025

Source: Author's chart using data from Housing Bureau, Hong Kong SAR Government. *Composite Waiting Time for Subsidised Rental Housing*. 2025. PDF file: *WT-Distribution-2025-ENG.pdf*.

During this prolonged waiting period, many households must remain in the private market, where rents are far higher. As shown in Table 5, private housing rents in 2025 averaged more than four times public rental housing rents per square meter. The gap was largest on Hong Kong Island, which contains some of Hong Kong's most central and expensive private housing markets. For applicants awaiting a public rental housing unit, time spent in the private market can impose a substantial and sustained financial burden rather than a short-term inconvenience.

Table 5. Public and Private Housing Distribution and Rent by Region, Hong Kong, 2025

District	Public housing household share	Private housing household share	Public rent (HK$/sqm IFA)	Private rent (HK$/sqm IFA)	Private/public rent ratio
Hong Kong Island	8.40%	21.75%	82	427	5.21x
Kowloon	34.20%	27.93%	92	380	4.13x
New Territories	57.40%	50.32%	74	290	3.92x
Total / average	100.00%	100.00%	80.83	344.94	4.27x

Note: Hong Kong's urban territory is commonly divided into Hong Kong Island, Kowloon, and the New Territories, which differ in density, housing composition, and proximity to major employment areas. Private-housing rent refers to the average monthly rent of flats under 70 sqm. Figures are as of March 31, 2025. Public rental-housing rent increases are subject to a 10% cap, while private-housing rents are not. The private/public rent ratio and total/average rent figures are the author's calculations.
Source: Author's table using data from Housing Bureau, Hong Kong SAR Government, Housing in Figures 2025.

This rent gap translates directly into rent burden. As shown in Table 6, public rental housing tenants face an estimated rent burden of approximately 13.9%. If these same households' income (HKD 20,500) were instead applied to the average private-market rent for a comparably sized unit, the implied rent burden would rise to 59.3%, a difference of more than 45 percentage points. This illustrates the financial exposure public housing tenants would face if they lost access to public housing, rather than an observed rent burden among households who currently rent privately.

Table 6. Monthly Rent and Estimated Rent Burden in Public and Private Housing, Hong Kong, 2025

Housing type	Monthly rent	Rent to income ratio
Public Rental Housing	HKD 2,849.19	13.90%
Private Housing	HKD 12,159.02	59.31%

Note: Figures are author's calculations. Estimated rent burden is calculated by combining average rent, floor area, household size, and income figures from official sources because Hong Kong does not report rent burden in the same format as New York City. For public rental housing, the calculation uses an average household size of 2.5 persons, interior floor area of 14.1 sqm per person, and monthly household income of HKD 20,500. The private-housing rent burden applies this same public-housing household income to the average private-market rent, representing a hypothetical burden rather than an observed rent burden among actual private-market renters.

Source: Author's table using data from Housing Bureau, Hong Kong SAR Government, Housing in Figures 2025; Census and Statistics Department, Hong Kong SAR Government, Hong Kong Annual Digest of Statistics 2025.

Together, Tables 5 and 6 show that the key challenge in Hong Kong's low-income housing system is whether households can access public rental housing before prolonged exposure to unaffordable private-market housing undermines their stability.

This access bottleneck reflects both persistent demand and uneven supply output. As Figure 7 shows, public rental housing completions fluctuated considerably over the past decade, ranging from 857 to 21,764 units, while the number of general applicants consistently remained above 100,000. This mismatch between uneven annual completions and sustained applicant demand helps explain why waiting times have remained persistently high.

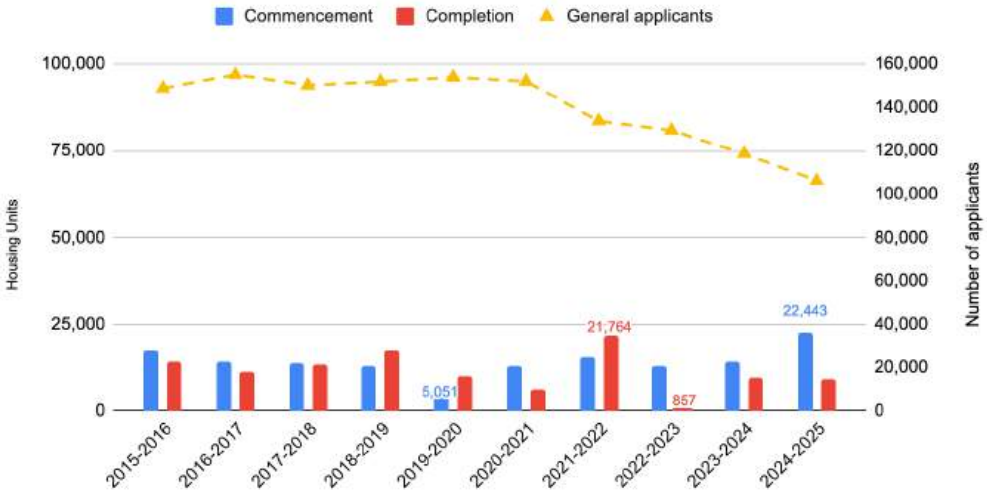


Figure 7. Public Rental Housing Commencements, Completions, and General Applicants, Hong Kong, 2015–2025

Source: Author's chart using data from Housing Bureau, Hong Kong SAR Government, *Housing in Figures 2025*; Census and Statistics Department, Hong Kong SAR Government, *Hong Kong Annual Digest of Statistics*, annual editions 2015–2025.

Rising construction costs may further constrain expansion. As Figure 8 shows, the cost of usable floor area in completed residential buildings roughly doubled between 2015-2016 and 2024-2025, from HK$62,015 to HK$125,527 per square meter. This increase may make it more difficult to expand public rental housing supply at the pace needed to reduce waiting times.

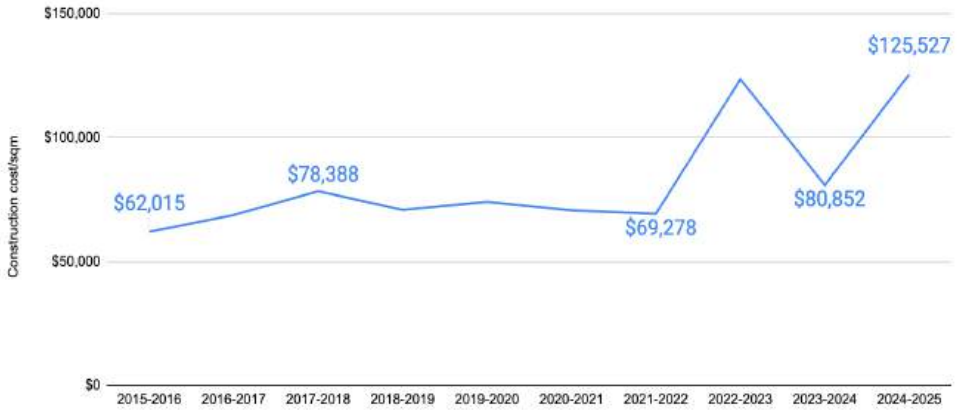


Figure 8. Construction Cost per Square Meter of Usable Floor Area in Residential Buildings, Hong Kong, 2015–2025

Note: Usable floor area includes residential floor area used for non-domestic purposes, such as clubhouses or recreational facilities, caretakers' offices or quarters, and transformer rooms.

Source: Author's chart using data from Census and Statistics Department, Hong Kong SAR Government, Hong Kong Annual Digest of Statistics 2025, p. 193.

Hong Kong thus illustrates a fundamental tension in large-scale public rental housing systems. Its dominant public rental housing program provides deep, durable affordability at scale. However, when supply pipelines are uneven and construction costs rise, the same system can produce a severe access bottleneck. The critical question is therefore not whether public rental housing provides protection, but how long households must wait before that protection begins.

4.2 Access to Opportunity

Hong Kong's territory is physically compact, but much of its land consists of steep hillsides and water, limiting the amount of developable space. Because of this compact geography, the distance between the New Territories and Hong Kong Island is more limited than comparable separations in larger metropolitan areas. At the same time, housing and employment are not evenly distributed across the city.

As shown in Figure 9, public rental housing households are more concentrated in the New Territories, which covers a much larger land area and accounts for 57.4% of public rental

housing households. Although Hong Kong's main business districts are concentrated around Central and Admiralty on Hong Kong Island,[8] the MTR and bus network provide broad cross-territory connectivity, reducing the degree of spatial isolation that might otherwise result from the concentration of public rental housing in the New Territories. The Transport Department reports mean journey times of 31 minutes for private vehicle/taxi trips and 45 minutes for public transport trips, with home-based work trips averaging the longest at 48 minutes.[9] These figures suggest that commute times in Hong Kong are generally moderate rather than extreme.

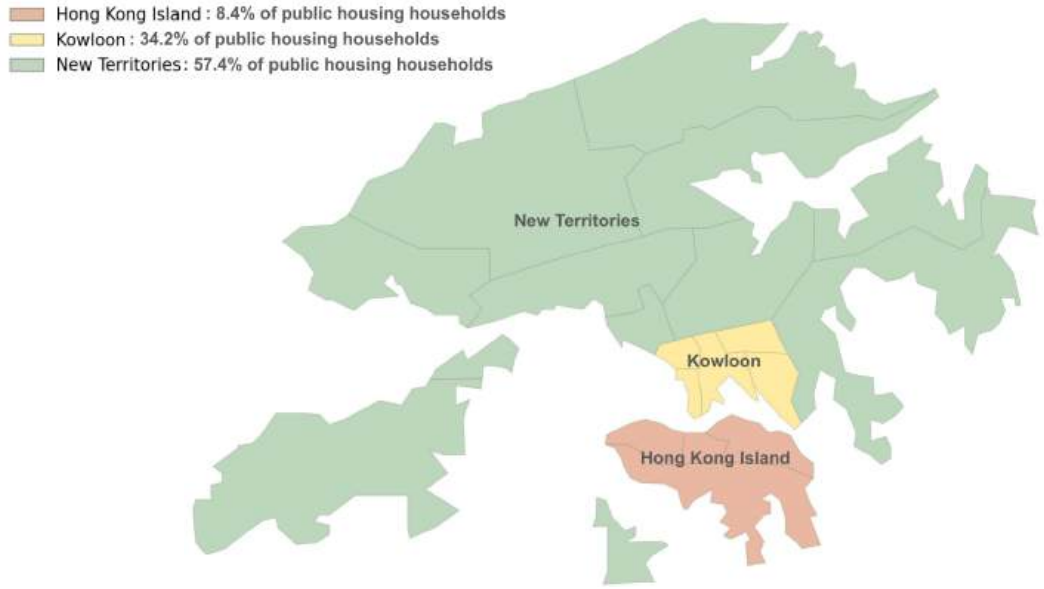


Figure 9. Public Housing Households by Region, Hong Kong, 2025

Note: Colors identify Hong Kong's three major geographic regions: Hong Kong Island, Kowloon, and the New Territories. Percentages show each region's share of public housing households, not intensity within the region.
Source: Author's map using SimpleMaps Hong Kong GIS boundary data (CC BY 4.0) and Housing Bureau, Hong Kong SAR Government, *Housing in Figures 2025*.

Overall, the analysis of housing stability and access to opportunity shows both the strength and limits of Hong Kong's heavily state-centered housing system. Public rental housing provides deep affordability once accessed, and the city's compact geography and transit network help keep commute times generally moderate. The main challenge for lower-income households appears to be less severe physical separation from jobs than delayed access to stable and affordable housing itself.

[8] Central and Admiralty are commonly treated as Hong Kong's core business districts because they contain major office, government, finance, and commercial functions.
[9] Transport Department, Hong Kong SAR Government, *Travel Characteristics Survey 2022 Final Report*, Figure 3.4.

5 Tokyo

5.1 Housing Stability

Tokyo addresses housing pressure through planning rules, redevelopment, and market-led housing production that allow the overall housing stock to expand. For low-income housing, this matters because a more elastic housing supply can reduce pressure on the rental market and prevent extreme scarcity. However, overall housing growth does not automatically guarantee deep affordability; the type of housing being added also matters.

Table 7 separates Tokyo's housing stock by construction period and tenure type. Local-government rental housing, including Toei Housing, functions as the main deeply affordable public housing option for low-income households. Urban Renaissance Agency (UR) and public-corporation rentals are quasi-public rental housing categories that serve a broader renter population and often operate closer to private-market rent levels. Issued houses refer to housing units owned or administered by companies, government agencies, or other organizations.

Table 7. Housing Stock by Construction Period and Tenure, Tokyo-to, 2023

Construction Year	Rented houses owned by local government	Rented houses owned by UR or public corporation	Rented houses owned privately	Issued houses	Owned houses
1970 or earlier	30,000	42,000	77,300	3,400	187,500
1971 to 1980	56,500	71,800	226,500	15,300	426,900
1981 to 1990	40,600	24,500	525,200	27,500	471,900
1991 to 2000	46,500	25,900	546,900	40,200	588,700
2001 to 2005	20,700	16,400	289,400	28,700	405,700
2006 to 2010	12,100	12,200	283,900	25,000	325,300
2011 to 2015	13,100	5,400	276,800	26,600	342,600
2016 to 2020	17,200	3,300	369,500	35,500	311,400
2021 to Sep. 2023	3,200	1,200	172,800	21,300	124,000

Source: Author's table using data from Statistics Bureau of Japan*, 2023 Housing and Land Survey, Basic Tabulation on Dwellings and Households,* Table 5-2.

Table 7 shows that privately owned rental housing is the largest category in nearly every post-1980 construction cohort, accounting for over 50% of all units shown in cohorts since 2016. By contrast, local-government and UR/public-corporation rental housing account for much smaller shares of recent construction. This distribution suggests that Tokyo's recent housing growth has been driven primarily by market-led rental and owner-occupied construction rather than by new public rental supply.

In the *2023 Housing and Land Survey*, the most recent available survey, Tokyo had about 8.20 million dwellings, including 896,500 vacant units, of which 629,000 were available for rent. Among the roughly 6.77 million principal households, only 444,000, or 6.55%, lived in public or quasi-public rental housing. These figures suggest that Tokyo's housing challenge is not simply a

shortage of total units, but the limited scale of deeply affordable public housing within a housing system dominated by private rental and owner-occupied units.

According to the survey, the average monthly rent of rented houses in Tokyo-to in 2023 was ¥88,266. As shown in Table 8, local-government rental housing provides the deepest affordability, while UR or public-corporation rental housing is priced close to the overall rental average and privately owned rental housing is slightly above it. This means that Tokyo's most affordable rental option is also relatively small in scale, leaving many lower-income households dependent on market-priced rental housing.

Table 8. Principal Households and Average Monthly Rent by Tenure, Tokyo-to, 2023

Tenure of Dwelling	Principal Households	Average Monthly Rent
Rented houses owned by local government	241,300	¥28,722
Rented houses owned by UR or public corporation	202,700	¥89,986
Rented houses owned privately	2,868,500	¥94,802
Issued houses	226,200	¥66,940
Owned houses	3,235,200	-

Note: — = not applicable.Average monthly rent does not control for unit size, housing quality, or number of rooms.
Source: Author's table using data from Statistics Bureau of Japan, *2023 Housing and Land Survey, Basic Tabulation on Dwellings and Households*, Table 126-2.

Households eligible for Tokyo public housing are generally those with certified monthly income of ¥158,000 or less, with a higher limit of ¥214,000 for some elderly, disabled, or single-parent households. Because this administrative threshold is based on certified monthly income rather than the annual income bands used in the Housing and Land Survey, this study uses an analytical income grouping instead: low-income households are defined as those with annual household incomes below ¥3 million,[10] lower-middle-income households as those earning ¥3 million to under ¥5 million, and middle- to upper-income households as those earning ¥5 million and above.

Table 9 shows that many lower-income households remain exposed to the more expensive segments of Tokyo's rental market. In the selected low- and lower-middle-income bands, rent burden in local-government rental housing remains below 20%, but local-government housing accounts for no more than about 15% of households in these bands. By contrast, private rental housing remains the dominant tenure, and households earning below ¥2 million face severe rent burdens above 50% in private rental housing.

Table 9. Estimated Rent Burden and Household Share in Selected Lower-Income Bands, Tokyo-to, 2023

[10] JKK Tokyo describes certified monthly income as total household income minus required deductions, divided by 12 months. The higher eligibility threshold of ¥214,000 per month equals ¥2,568,000 per year before adding applicable deductions. Since deductions vary by household composition, this paper uses the Housing and Land Survey's "under ¥3 million" annual income band as an approximate analytical category rather than an official eligibility cutoff.

Annual Income band	Local-government rental burden	UR/public corporation rental burden	Private rental burden	Local-government rental share	UR/public corporation rental share	Private rental share
1 to under 1.5 million yen	19.48%	61.70%	66.39%	14.79%	4.19%	45.70%
1.5 to under 2 million yen	16.88%	47.86%	51.25%	10.76%	4.52%	46.96%
2 to under 3 million yen	14.05%	36.58%	36.70%	5.89%	3.62%	52.61%
3 to under 4 million yen	14.07%	28.96%	28.22%	3.32%	3.46%	51.39%
4 to under 5 million yen	13.48%	24.67%	23.77%	2.26%	2.88%	45.85%

Note: Figures are author's calculations. Estimated rent burden is calculated by annualizing average monthly rent and dividing it by the midpoint of each annual income band. Household shares are calculated from Housing and Land Survey household counts. Figures are approximate because income is reported in ranges rather than exact amounts.
Source: Author's calculation using data from Statistics Bureau of Japan, 2023 Housing and Land Survey: Basic Tabulation on Dwellings and HouseholdsJapan, Prefecture, Shi, Ku, Machi and Mura. Table 126-2.

Under the Public Housing Act and the Tokyo Metropolitan Government Public Housing Ordinance, Toei Housing provides low-rent public housing as a safety net for low-income households that have difficulty securing housing in the private market. Since Tokyo does not report public-housing need primarily through a single average waiting-time measure, this paper assesses access delay through application ratios, lottery results, and point-system magnification tables, as shown in Table 10.

Table 10. Toei Housing Application Ratios by Recruitment Type, Tokyo, 2025

Recruitment type	Housing category	Units offered	Applicants	Application ratio
Lottery	Single-person housing	155	7,515	48.5×
Lottery	Single-person wheelchair-accessible housing	4	29	7.3×
Lottery	Silver Peer, single-person senior housing	30	2,446	81.5×
Lottery	Silver Peer, two-person senior housing	8	272	34.0×
Lottery	Single-person housing where death occurred inside unit	93	2,500	26.9x
Lottery	Two-or-more-person household housing where death occurred inside unit	7	419	59.9x
Point system	Family-oriented general housing	664	3,808	5.7×
Point system	Wheelchair-user household housing	52	32	0.6×

Source: Author's table using data from JKK Tokyo, August 2025 Toei Housing recruitment lottery and point-system magnification tables

Table 10 shows that access to Toei Housing remains highly competitive, especially for single-person and senior housing. Single-person Silver Peer senior housing had an 81.5-to-1 application ratio, while general single-person housing had a 48.5-to-1 ratio. Point-system categories are allocated through need-based selection rather than ordinary lottery, but family-oriented units also remain limited, with 3,808 applicants for 664 units. These figures

suggest substantial excess demand and show that Tokyo's deeply affordable public rental sector remains difficult to access even within a broader housing system that is more supply-responsive. Although Tokyo's broader housing system is relatively supply-responsive, that responsiveness operates mainly through market housing rather than through the expansion of deeply affordable public rental units. Private rental and owner-occupied construction can expand the total housing stock, but these units remain largely market-priced and therefore cannot substitute for low-rent public housing. As a result, Tokyo's flexible housing production may reduce severe overall scarcity, but it does not by itself guarantee deep affordability for low-income households. The system's main limitation is not that it fails to produce housing, but that much of the housing it produces is not affordable enough for the lowest-income households.

5.2 Access to Opportunity

Tokyo's transit network resembles a spiderweb: radial rail corridors connect suburban areas to the core, while circular routes link those corridors to one another. Its central business core is the Marunouchi–Otemachi–Yurakucho district around Tokyo Station, while major subcenters include Shinjuku, Shibuya, and Ikebukuro. These employment centers are distributed across the network rather than concentrated in a single downtown area. This structure allows households outside the central core to maintain relatively efficient access to multiple employment destinations. As shown in Table 10, over half of households in the Tokyo metropolitan area live within 1 kilometer of a train station, while close to 85% live within 2 kilometers. These figures suggest relatively high rail accessibility and are consistent with Tokyo's transit-oriented urban structure.

Table 11. Household Distribution by Distance to Nearest Train Station, Tokyo Metropolitan Area, 2023

Distance to nearest train station	Share of households
Under 500m	25.44%
500 to under 1000m	31.67%
1000 to under 2000m	27.23%
2000m and over	15.66%

Note: "Tokyo metropolitan area" refers to the official Kanto Major Metropolitan Area used in the Housing and Land Survey. Bus-stop subcategories in the original survey are omitted; households are grouped only by distance to the nearest train station.
Source: Author's table using data from Statistics Bureau of Japan, *2023 Housing and Land Survey, Basic Tabulation on Dwellings and Households, Major Metropolitan Areas and Metropolitan Areas,* Table 206.

This rail-access distribution helps explain why access to opportunity in Tokyo depends heavily on station proximity and rail connectivity to multiple employment centers, rather than simply on distance from the Marunouchi–Otemachi–Yurakucho central business core.

Table 12. Median Commuting Time by Housing Type, Tokyo-to, 2023

Housing Type	Median Commuting Time (Minutes)

Rented houses owned by local government	31.6
Rented houses owned by UR or public corporation	47.5
Rented houses owned privately	39.9
Issued houses	36.0

Source: Author's table using data from Statistics Bureau of Japan, *2023 Housing and Land Survey, Basic Tabulation on Dwellings and Households: Japan, Prefecture, Shi, Ku, Machi and Mura*, Table 127-1-2.

Table 12 shows that the median commuting time for local-government rental housing residents was 31.6 minutes, shorter than that for private rental housing residents at 39.9 minutes and UR or public-corporation rental housing residents at 47.5 minutes. This suggests that access to opportunity in Tokyo depends more on how housing is positioned within the wider rail network and metropolitan development structure than on a simple trade-off between cheaper rent and longer commute time.

Tokyo's access-to-opportunity picture is stronger than its housing-stability picture. The rail network reduces commuting penalties for many lower-income households, and the dispersion of employment centers means that living outside the central business core does not necessarily imply poor employment access. The main limitation is affordability: many lower-income households can reach jobs within moderate commute times, but still face high rent burdens because the deeply affordable public rental housing sector is too small to serve most of them.

6 Limitations

This study has several limitations. First, the three cities do not report housing data in identical categories, so the comparison relies on functionally equivalent indicators rather than perfectly identical measures. Second, rent measures do not fully control for unit size, housing quality, household composition, or neighborhood amenities. Third, the income categories used for Tokyo and Hong Kong do not perfectly match New York City's rent-burden categories. Fourth, the transit-access measures differ across cities: New York City uses author-calculated travel-time bands, Hong Kong uses official journey-time statistics and regional public-housing distribution, and Tokyo uses station-distance and commute-time-by-tenure indicators. Finally, the analysis is descriptive and comparative rather than causal; it identifies patterns between housing systems, affordability, delay, and access, but does not isolate the independent causal effect of each policy.

7 Conclusion

This comparison of New York City, Hong Kong, and Tokyo shows that low-income housing outcomes depend not only on the amount of government intervention, but on how each city combines housing supply, affordability support, and transit access to employment. Across all three cases, low-income households face the greatest instability when they are pushed into the private rental market. However, the reason they remain in that market differs by city.

In Hong Kong, the main problem is waiting time: public rental housing provides deep affordability, but many households must first spend years in the private market before receiving that protection. In Tokyo, the main problem is scale. The city supports broad housing production and maintains strong rail access, but its deeply affordable public housing sector remains too small to serve most lower-income renters. In New York City, both problems appear together. Multiple housing programs reduce instability for some households, but long waits, limited availability, and a shrinking regulated stock leave many households exposed to the private market. At the same time, many of the lowest-rent areas are farther from major employment centers, creating an additional trade-off between affordability and access to jobs.

The comparison also shows that affordability cannot be evaluated separately from access to employment. Hong Kong's compact geography and strong transit network help reduce commute penalties for public housing residents. Tokyo's rail-oriented structure allows lower-cost public housing to remain relatively connected to jobs. New York City's more Manhattan-centered transit network creates a clearer trade-off between lower rent and shorter commute times for many outer-borough residents. For policymakers, the comparison suggests that affordable housing should be planned as part of a broader urban system, not only as a rent subsidy or unit-production target. Housing programs are likely to be more effective when affordability, timely access, and transit connection to employment are considered together.

Works Cited

Census and Statistics Department, Hong Kong SAR Government. *Hong Kong Annual Digest of Statistics 2025*. Hong Kong: Hong Kong SAR Government, 2025.

Census and Statistics Department, Hong Kong SAR Government. *Hong Kong Annual Digest of Statistics*. Annual editions, 2015–2025. Hong Kong: Hong Kong SAR Government.

Housing Bureau, Hong Kong SAR Government. *Composite Waiting Time for Subsidised Rental Housing*. Hong Kong: Hong Kong SAR Government, 2025.

Housing Bureau, Hong Kong SAR Government. *Housing in Figures 2025*. Hong Kong: Hong *Kong* SAR Government, 2025.

Japan Ministry of Land, Infrastructure, Transport and Tourism. *Urban Planning System in Japan / Land Use Zones*. Government of Japan.

JKK Tokyo. *August 2025 Toei Housing Recruitment Lottery and Point-System Magnification Tables*. Tokyo: Tokyo Metropolitan Housing Supply Corporation, 2025.

Metropolitan Transportation Authority. *GTFS Subway and Bus Feeds*. New York: Metropolitan *Transportation* Authority, 2023.

New York City Department of Housing Preservation and Development. *New York City Housing and Vacancy Survey: Selected Initial Findings*. 2011, 2014, 2017, 2021, and 2023.

New York City Department of Housing Preservation and Development. "New York City *Announces* Back-to-Back Record Years for Creating and Connecting New Yorkers to Affordable Housing." July 29, 2024. https://www.nyc.gov/site/hpd/news/032-26/new-york-city-back-to-back-record-years-creating-connecting-new-yorkers-to

New York City Housing Authority. *2024–2028 Capital Plan Narrative*. 2024. *https*://www.nyc.gov/assets/nycha/downloads/pdf/2024-2028-Capital-Plan-Narrative.pdf

New York City Housing Authority. *Draft Public Housing Agency Plan: Draft Annual Agency Plan for Fiscal Year 2026*. New York: New York City Housing Authority, 2025.

New York City Mayor's Office of Operations. *Preliminary Mayor's Management Report, Fiscal Year 2026: Human Resources Administration*. New York: City of New York, 2026. https://www.nyc.gov/assets/operations/downloads/pdf/pmmr2026/hra.pdf

Office of the New York City Comptroller Brad Lander. *Evictions Up, Representation Down: How New York City Is Undermining the Right to Counsel*. New York: Office of the New York City Comptroller, May 2025. https://comptroller.nyc.gov/wp-content/uploads/documents/Evictions-Up-Representation-Down.pdf

r5py contributors. (2024). r5py: Rapid realistic routing with R5 in Python (Version 1.1.3) *[*Software]. https://r5py.readthedocs.io/stable/. https://doi.org/10.5281/zenodo.7060437

SimpleMaps. *Hong Kong GIS Boundary Data*. SimpleMaps, CC BY 4.0.

Statistics Bureau of Japan. *2023 Housing and Land Survey: Basic Tabulation on Dwellings and Households*. Government of Japan, 2023.

Statistics Bureau of Japan. *2023 Housing and Land Survey: Basic Tabulation on Dwellings and Households, Major Metropolitan Areas and Metropolitan Areas*. Government of Japan, 2023.

Tokyo Metropolitan Government. *Tokyo Metropolitan Government Public Housing Ordinance*.

Transport Department, Hong Kong SAR Government. *Travel Characteristics Survey 2022 Final Report*. Hong Kong: Hong Kong SAR Government, 2022.

U.S. Census Bureau. *2023 TIGER/Line Shapefiles*. U.S. Department of Commerce, 2023.

U.S. Census Bureau. *American Community Survey 2023 5-Year Estimates, Table B25003: Tenure*. U.S. Department of Commerce, 2023.

U.S. Census Bureau. *American Community Survey 2023 5-Year Estimates, Table B25064: Median Gross Rent*. U.S. Department of Commerce, 2023.

U.S. Census Bureau. *American Community Survey 2023 5-Year Estimates, Table B25065: Aggregate Gross Rent*. U.S. Department of Commerce, 2023.

Mental Toughness in Adolescents By Olivia Shekou

Abstract

Mental toughness, characterized by resilience and a proactive drive to challenge oneself for growth, plays a critical role in the academic performance of high school students. While not unique to adolescents, high school is a key developmental period for cultivating these skills and shaping future success. This review explores the key factors that influence the development of mental toughness and examines how it relates to academic achievement and mental health. It also evaluates whether mental toughness is primarily genetically inherited or shaped by environmental influences. Given its importance during adolescence, these findings raise important questions about whether high schools should play a more active role in fostering mental toughness and integrating it into their curricula.

Introduction

The prevalence of mental illness among children and adolescents in the United States is a growing public health concern. According to the CDC's Youth Risk Behavior Survey Data Summary & Trends Report: 2013-2023, 40% of U.S high school students reported persistent feelings of sadness or hopelessness. Additionally, the CDC also finds that 20% of students considered attempting suicide, and 9% reported suicide attempts ("Adolescent and School Health").

These rates have increased significantly since the onset of the COVID-19 pandemic. During the first year of the pandemic, the global prevalence of depression increased substantially, in part due to the absence of "school connectedness" during a time of severe disruption. Fewer than half of youth reported feeling close to others at school, and students who did not feel connected to school reported markedly higher rates of persistent sadness or hopelessness than those who did (53% vs. 35%), and roughly double the rate of attempted suicide (12% vs. 6%) ("New CDC data illuminate youth mental health threats during the COVID-19 pandemic").

Untreated mental health challenges among youth have significant academic, social, and economic ramifications. Mental health symptoms have been linked to declining academic performance, increased truancy, and higher dropout rates. At the same time, many students experienced significant learning loss during this period, widening achievement gaps and leaving them underprepared for higher education and an evolving workforce. As a result, colleges and universities are facing increased demand for mental health services while also working to support students with gaps in academic readiness (Pappano). How these challenges are addressed or neglected will not only shape the future of this generation but also the broader social and economic outcomes of our nation.

The drivers of this youth mental health crisis are multifactorial. Factors associated with mental health symptoms in adolescents include lack of social support, lower socio-economic

status (SES), familial history of mental health symptoms, school bullying, and substance use. One notable correlation is increased academic pressure, largely due to global academic competition, especially with China and South Korea, which are widely considered leaders in STEM.

The U.S. education system's emphasis on grades, test scores, and short-term outcomes often comes at the expense of the student's long-term growth, adaptability, and resilience. Post-pandemic conditions, including teacher shortages and larger class sizes, have further reduced individualized support.

With the rapid expansion of artificial intelligence (AI), workforce demands increasingly prioritize adaptability and problem-solving skills. Many schools fail to equip students with real-world experiences and transferable skills necessary for success in the modern workforce. As AI becomes central to the job market, younger generations must develop strong adaptability and problem-solving abilities. "AI's recent rapid growth is unlikely to slow down in the near future. In fact, global AI adoption by organizations is set to expand at a CAGR [compound annual growth rate] of 35.9% between 2025 and 2030" (Howarth). This rapid growth presents new challenges, requiring employees to think critically, adapt quickly, and effectively integrate AI into their professions. However, the current middle and high school education system relies on outdated, memorization-based approaches and places insufficient emphasis on developing mental skills and transferable competencies. In this evolving economic landscape, many studies argue that the U.S. education system must prioritize qualities such as adaptability and resilience.

One such quality is mental toughness. Mental toughness is defined as a mindset of perseverance, adaptability, problem-solving, and grit. It is often referred to as the four "C's" model of mental toughness: control, commitment, challenge, and confidence (Clough et al.). Although often confused with resilience, mental toughness extends beyond resilience in its proactive tendency to seek out challenges as opportunities for growth. Studies by Zappala-Piemme et al. (2023) posit that the most effective way to cultivate mental toughness in school settings is to emphasize skills such as stress management, perseverance, coping mechanisms, and creative problem-solving. This study supports the assertion that mental toughness should be structurally integrated into middle and high school curricula, rather than treated as a supplementary concept. However, the U.S. education system often overlooks this critical developmental window, during which stress levels are particularly high. Instead of prioritizing adaptability and growth, it frequently emphasizes perfection and performance through rigid metrics and ranking systems (Zappala-Piemme et al.). As a result, many students become hesitant to take risks or embrace challenges, experiences that are essential to building resilience and developing mental toughness. Although high school represents a critical opportunity to develop these skills, they are rarely prioritized within the curriculum, potentially compromising students' long-term success and well-being.

The goal of this paper is to examine how mental toughness affects academic performance and mental health in high school students. This review also seeks to identify effective strategies for fostering mental toughness within educational settings, as well as the barriers within schools that hinder its development.

Methods

This review was conducted using Google Scholar to locate articles in the scope of the current topic. The goal of this method was to analyze information within the specific adolescent population and to further define the scope of the research topic. The search terms used in Google Scholar were: Mental Toughness IN high school students AND academic performance AND mental health AND outcomes. Google Scholar automatically converts these search terms into multiple results, capturing articles related to mental toughness and its relationship to academic performance, mental health, and long-term outcomes in adolescents. The search was limited to studies published between 2000 and 2025. Only articles published in English focusing on adolescents (primarily high school students aged 14-18) were included. Because the available evidence base on adolescent mental toughness is concentrated outside the United States, studies conducted in comparable secondary-education systems (e.g., the United Kingdom, Switzerland, and Turkey) were included and are treated as broadly relevant to U.S. adolescents while acknowledging cross-cultural differences. All included studies examined the relationship between mental toughness and academic and mental health outcomes and used either quantitative or qualitative research data. Studies were excluded if they focused on elementary, middle school, or exclusively college populations. Opinion-based studies were also excluded. After screening these articles for eligibility, six studies met the inclusion criteria and were selected for analysis. Measures of mental toughness, academic performance outcomes, and mental health variables were analyzed from each study to better understand how mental toughness influences academic and psychological outcomes in adolescents.

Results

See a summary of findings in Table 1.

Study and Design	**Sample Characteristics**	**Independent Variable(s)**	**Dependent Variable(s)**	**Summary of Study Findings**
Sağkal and Özdemir (2019); Cross-sectional	Turkish adolescents (ages 14–18)	Parenting style (strength-based parenting)	Mental Toughness (psychological outcomes)	Strength-based parenting was directly and indirectly (via mental toughness) linked to lower psychological distress

survey				and higher subjective happiness in adolescents.
Gerber et al. (2013) Cross-sectional survey	Swiss high school and young adult students	Environment and age	Mental Toughness; stress resilience	This study demonstrated how mental toughness increases with age and is shaped by developmental and contextual factors.
McGeown et al. (2018) Cross-sectional Study	U.S. adolescents	Environment and Psychological traits	Mental Toughness	An MT scale is created and correlates with improved academic performance, coping in schools, and mental health.
McGeown et al. (2017) Cross- sectional Study	UK secondary school students	Environment and Psychological traits	Mental Toughness	An MT scale is created and correlates with improved academic performance, coping in schools, and mental health.
Lin et al. (2017) Systemic review	Students and adults	Mental Toughness	Academic/ Workforce performance; wellbeing	MT is correlated with increased academic performance and higher future income, supporting its relevance to long-term success.
St. Clair-Thompson et al. (2014)	UK adolescents (ages 11–16)	Mental Toughness (4C's / MTQ48)	Academic attainment, attendance, classroom behavior,	Findings from this study illustrate how mental toughness reduces stress and depressive symptoms

			peer relationships	and becomes a more effective resilience factor with age. Aspects of mental toughness (particularly control of life) were significantly associated with academic attainment and attendance; confidence in abilities and interpersonal confidence were linked to peer relationships, supporting mental toughness as a useful construct in education.

Table 1: Summary of Findings

Discussion

Mental toughness is influenced by a variety of factors, both environmental and individual-level. These predictors can be immediate (e.g., family and parenting), broader (e.g., educational and community contexts), and intrinsic (e.g., personality and age). Overall, these studies (See Table 1, #1-4) indicate that adolescent mental toughness is shaped by complex interactions between these influences. Below, the results are discussed in further detail.

Environmental and Community Level Predictors

Parenting and family environments play a crucial role in fostering mental toughness from an early age. In 2019, Sağkal and Özdemir conducted a cross-sectional survey of United States adolescents in order to analyze the outcomes of strength-based parenting. This form of parenting is centered on encouragement and recognition of a child's natural strengths in order to, according to Dr. Lea Waters, "cultivate positive states, processes and qualities in one's children" (2). Results demonstrated that strength-based parenting decreased behavioral problems and increased perseverance and resilience. Thus, supportive family environments are key predictors of mental toughness.

Likewise, school and community environments play a key role in improving mental toughness and academic engagement. According to Dr. Sara McGeowan, a supportive

educational environment is associated with increased mental toughness and overall achievement. The 4C's model of mental toughness (commitment, challenge, control, and confidence) is directly related to one's school environment ("The Development and Validation of a Mental Toughness Scale for Adolescents"). Commitment and perseverance in completing challenging assignments and overcoming obstacles are correlated with an adolescent's ability to see new situations as opportunities for development. The adolescent's willingness to engage with opportunities for self-development was influenced by their confidence, past success experiences and teacher support, highlighting the key environmental/non-cognitive factors in shaping mental toughness McGeown et al.'s work also illustrates the long-term impact of this: "those who were confident in their abilities would feel comfortable attempting new or difficult tasks, whereas those with high levels of interpersonal confidence would feel confident in social situations, particularly in new or unfamiliar environments [in the future]" (1). Thus, the school environment has been found to play a key role in fostering Mental Toughness.

At a broader level, community-level factors also impact the development of Mental Toughness. Factors such as community size impact mental toughness. Gerber et al. (2012) compared high school and undergraduate students in different community settings, examining how community-level factors (the size of the town/school environment) correlated with stress exposure, social support, and opportunities for developing coping skills. Results implied that close-knit communities provided more accessible support while larger communities offered more diverse experiences and challenges, both of which fostered the development of coping skills in adolescents (Gerber et al.). Therefore, both types of environments allowed for the development of adaptive strategies/skills that fostered mental toughness, although in different ways.

Overall, these findings indicate that one's environment (nurturing family and school communities) cultivates mental toughness in adolescents by arming them with confidence, skills, and a successful mindset.

Individual-Level factors

Finally, at the individual level, evidence from Lin et al. (2017) and St. Clair-Thompson et al. (2014) showed that age, personality, and pre-existing psychological traits also contribute to mental toughness. Lin et al. (2017) illustrated how mental toughness is portrayed through confidence and resilience. Evidence from St. Clair-Thompson et al. (2014) highlights that specific psychological traits, as well as older age, predict higher mental toughness. Findings from the study suggest that students high in mental toughness are also those who are able to reach a state of flow (i.e., a state of effortless engagement). Additionally, they found that people who had psychological traits such as self-efficacy, optimism, and motivation had higher levels of mental toughness.

Implications

Studies suggest that adolescent mental toughness is associated with a variety of beneficial outcomes, including psychological, academic, and long-term performance-related outcomes (See Table 1, #5-6). These benefits are discussed in detail below.

Psychological Outcomes

Research illustrated how adolescents with higher mental toughness exhibited fewer signs of perceived stress and depressive symptoms (Gerber et al.). This highlights mental toughness as a protective resource. Additionally, strength-based parenting, which is correlated to mental toughness and lower psychological distress, sets the stage for individuals' development and success down the road in educational and workforce settings (Sağkal and Özdemir). Strength-based parenting allows adolescents to face setbacks while actively engaging in learning opportunities.

Academic Outcomes

Academic outcomes are positively influenced by mental toughness as well. McGeown et al. (2016) found that adolescents with higher mental toughness exhibited confidence in both academic and social situations, priming them for future success. Similarly, St. Clair-Thompson et al. (2014) found that mentally tougher students actively seek out opportunities for personal growth, furthering their academic success. Together, these studies suggest that mental toughness is linked to both greater confidence and growth-motivation.

Community-Level Outcomes

Mental toughness fosters the development of coping strategies and adaptive skills that allow adolescents in both close-knit and larger communities to engage in learning opportunities and overcome challenges (Gerber et al.). This highlights the effects of mental toughness beyond just a classroom, but as a means of priming an adolescent to be a successful and beneficial contributor to a community.

Long-term and Workforce Outcomes

Long-term and workforce outcomes are strongly interwoven with mental toughness. The systematic review by Lin et al. (2017) highlights that mental toughness is strongly associated with academic performance in adolescents and successful outcomes in the workplace. Adolescents with higher mental toughness are more likely to achieve higher educational attainment, which translates into higher income, increased job performance, and career opportunities. Mental toughness allows them to develop problem-solving abilities from a young age that they carry into professional environments. Additionally, this review highlighted how mental toughness is not a fixed quality, but one that can be fostered and enhanced by school settings, positive feedback, and other interventions during adolescence.

Conclusion

Mental Toughness is a multidimensional trait linked to positive outcomes in academic performance, mental health, and long-term success. It is shaped by a combination of environmental influences and individual factors over time. Those with greater mental toughness are proven to be more successful in many aspects of life. Despite these benefits, it is still unclear whether mental toughness can ensure similar benefits for individuals from lower socioeconomic backgrounds or conditions where external obstacles limit opportunities for success. Regardless, mental toughness is still an invaluable resource that should be taught from an early age to foster achievement on both an individual level and as a society.

Works Cited

"Adolescent and School Health." *CDC*, Centers for Disease Control, 29 Nov. 2024, www.cdc.gov/healthy-youth/mental-health/. Accessed 1 Aug. 2026.

Clough, Peter, et al. "Mental Toughness: The Concept and Its Measurement." *Solutions in Sport Psychology*, edited by Ian Cockerill, Thomson Learning, 2002, pp. 32–43, www.researchgate.net/publication/313119986_Mental_toughness_The_concept_and_its_measurement. Accessed 13 Mar. 2026.

Gerber, Markus, et al. "Are Adolescents With High Mental Toughness Levels More Resilient Against Stress?" *Stress and Health*, vol. 29, 2013, pp. 164-71, dpl6hyzg28thp.cloudfront.net/media/Stress_and_Health_-_2012_-_Gerber_-_Are_Adolescents_With_High_Mental_Toughness_Levels_Mo_cxTuOfS.pdf. Accessed 1 Aug. 2026.

Howarth, Josh. "45+ NEW Artificial Intelligence Statistics (January 2026)." *Exploding Topics*, 27 Mar. 2026, explodingtopics.com/blog/ai-statistics#. Accessed 1 Aug. 2026.

Lin, Ying, et al. "Mental Toughness and Individual Differences in Learning, Educational and Work Performance, Psychological Well-being, and Personality: A Systematic Review." *Frontiers in Psychology*, vol. 8, 11 Aug. 2017, https://doi.org/10.3389/fpsyg.2017.01345.

McGeown, Sarah, et al. "The Development and Validation of a Mental Toughness Scale for Adolescents." *Journal of Psychoeducational Assessment*, vol. 36, no. 2, Mar. 2018, pp. 148–61. *SAGE Journals*, https://doi.org/10.1177/0734282916673512. Accessed 1 Aug. 2026.

McGeown, Sarah, et al. "Understanding and Supporting Adolescents' Mental Toughness in an Education Context." *Psychology in the Schools*, vol. 54, no. 2, Feb. 2017, pp. 196–209. *Wiley Online Library*, wiley.com. Accessed 1 Aug. 2026.

"New CDC data illuminate youth mental health threats during the COVID-19 pandemic." *CDC*, Centers for Disease Control, 2022, archive.cdc.gov/#/details?url=https://www.cdc.gov/media/releases/2022/p0331-youth-mental-health-covid-19.html. Accessed 18 Jan. 2026.

Pappano, Laura. "Colleges Revamp Career Services to Address Gen Z's Work-Life Expectations." *Washington Post*, 20 May 2025, www.washingtonpost.com/education/2025/05/20/college-career-services-gen-z-work-life-balance/. Accessed 1 Aug. 2026.

Sağkal, Ali Serdar, and Yalçın Özdemir. "Strength-based Parenting and Adolescents' Psychological Outcomes: The Role of Mental Toughness." *Journal of Psychologists and Counsellors in Schools*, vol. 29, no. 2, Dec. 2019, pp. 177-89, https://doi.org/10.1017/jgc.2019.2.

St Clair-Thompson, Helen, et al. "Mental Toughness in Education: Exploring Relationships with Attainment, Attendance, Behaviour and Peer Relationships." *Educational Psychology*, vol. 35, no. 7, 11 Mar. 2014, pp. 886-907, https://doi.org/10.1080/01443410.2014.895294.

---. "Mental Toughness in Education: Exploring Relationships with Attainment, Attendance, Behaviour and Peer Relationships." *Educational Psychology*, vol. 35, no. 7, 11 Mar. 2014, pp. 886-907, https://doi.org/10.1080/01443410.2014.895294.

"YRBS Data Summary & Trends Report." *CDC*, Centers for Disease Control, 6 Aug. 2024, www.cdc.gov/yrbs/dstr/index.html.

Zappala‐Piemme, Kerri E., et al. "Building Mental Toughness: A Middle School Intervention to Increase Grit, Locus of Control, and Academic Performance." *Psychology in the Schools*, vol. 60, no. 8, 23 Mar. 2023, pp. 2975-90, https://doi.org/10.1002/pits.22893.

Treatment of Alzheimer's disease using Stem Cells and CRISPR-based genome editing systems By Teruki Otake

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the accumulation of Amyloid Beta (Aβ) 42, hyperphosphorylated tau, and chronic neuroinflammation. Current therapies provide symptomatic relief but do not halt or cure the underlying pathological processes. Advances in stem cell biology and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein (Cas)- based gene editing have enabled new approaches to investigating AD mechanisms and developing potential regenerative therapies. This review summarizes recent work using neural stem cells (NSCs), mesenchymal stem cells (MSCs), and induced pluripotent stem cells (iPSCs), as well as CRISPR platforms such as Cas9, Cas12a, Cas12f, and CasΦ. CRISPR-edited iPSCs and iPSC-derived NSCs allow precise modification of genes implicated in AD, including APP, PSEN1, PSEN2, and APOE4. These engineered models replicate hallmark features such as Aβ42 accumulation, tau abnormalities, and age related decline in stem cell function, providing valuable human-based systems for mechanistic studies and therapeutic screening. In parallel, transplantation studies demonstrate that NSCs and MSCs can reduce Aβ and tau burden, enhance neurogenesis and synaptic plasticity, and partially improve cognitive outcomes. These benefits may arise from direct neuronal replacement as well as paracrine secretion of neuroprotective and anti-inflammatory factors. CRISPR-mediated deletion of human leukocyte antigen (HLA) genes and upregulation of HLA E and HLA G offer strategies to improve immune compatibility of transplanted cells, although challenges remain, including off-target effects, incomplete immune protection, and risks of uncontrolled proliferation. Ethical considerations are particularly important for germline editing of APOE4. Continued refinement of gene-editing tools, rigorous safety evaluation, and responsible regulatory oversight will be essential to translating these technologies into effective therapies for AD.

Introduction

Alzheimer's disease (AD) was first reported in 1906 by the German psychiatrist Alois Alzheimer, who described a patient with progressive memory loss together with characteristic neuropathology, including extracellular plaques and intracellular neurofibrillary changes (Thakor et al.). Research indicates that AD arises from multiple interacting pathological processes, such as the accumulation of Aβ and the hyperphosphorylation of tau protein (Bermudez et al.). The accumulation of Aβ is driven by the abnormal cleavage of amyloid precursor protein by β-secretase and γ-secretase (Bahnassawy et al.). Mutations in Presenilin1 and Presenilin2, which encode components of γ-secretase, increase the relative production of the Aβ42 isoform that is particularly prone to aggregation. Excess extracellular Aβ aggregates form plaques that induce neurotoxicity through mechanisms involving synaptic dysfunction, oxidative stress, and neuroinflammation (Bermudez et al.).

Hyperphosphorylation of the tau protein also plays a central role in neuronal degeneration. Increases in intracellular Ca2 activate kinases such as GSK3β and CDK5, which drive abnormal tau phosphorylation and lead to the formation of neurofibrillary tangles (Ge et al.). In addition to these mechanisms, Apolipoprotein E contributes to Aβ42 clearance。 APOE binds with Aβ42 to remove it from the brain (H. Liu et al.). People who have APOE4 are more likely to develop AD than those who have APOE3 because proteins synthesized from APOE3 bind with Aβ42 more strongly than those synthesized from APOE4 (H. Liu et al.).

Despite more than a century of scientific progress, AD remains incurable (Breijyeh and Karaman et al.). Current drug treatments provide symptomatic relief but do not slow or stop the underlying disease processes that involve Aβ42 accumulation, tau hyperphosphorylation, and chronic inflammation. As a result, AD continues progressing in affected individuals despite the use of available medical interventions. The global burden of the disease highlights the urgency of this challenge. Worldwide dementia cases were estimated at 57.4 million in 2019, with a ninety-five percent uncertainty interval of 50.4 to 65.1 million. This number is projected to rise to 152.8 million by 2050 with a ninety-five percent uncertainty interval of 130.8 to 175.9 million (Arayici and Kose et al.). Current technologies, such as CRISPR-Cas-based systems and stem cells are enabling us to unravel the mechanism of AD and seek enhanced approaches to AD therapy.

Taken together, these observations highlight the need for an integrated evaluation of emerging stem cell and CRISPR-based approaches. This review, therefore, examines current genome editing platforms, stem cell types relevant to AD research, experimental disease models derived from these technologies, and their potential therapeutic applications and limitations.

1. CRISPR-Cas system

Precise genome manipulation has become central to both mechanistic studies of AD and the development of regenerative strategies. To understand how these applications are achieved, it is essential to outline the molecular basis of CRISPR-mediated gene editing.

1.1 CRIPSR-Cas9

CRISPR Cas9 is revolutionary for its cost-effectiveness and high specificity, enabling accurate and efficient gene editing. CRISPR and CRISPR-associated proteins originate from prokaryotes and function as an immune system to defend against invading viruses (Figure 1) (Adli).

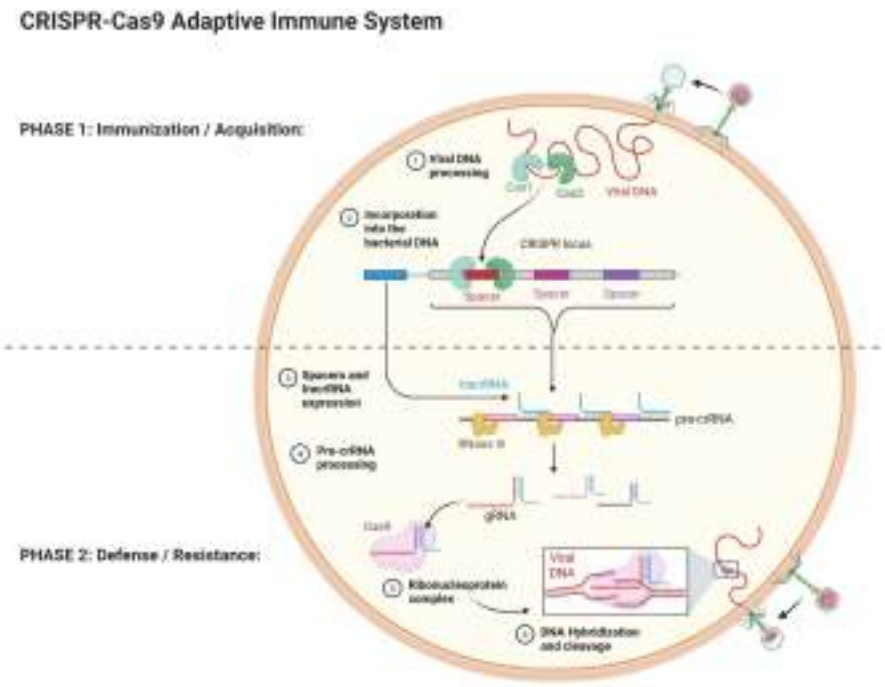


Figure 1. The CRISPR-Cas system found in Bacterial Cells.

Cas1 and Cas2 cut DNA sequences derived from invading viruses and process them into protospacers that are integrated into the CRISPR array as spacers (Adli et al.; Killelea et al.). RNA polymerase transcribes the spacers and repeats to produce pre-CRISPR RNA. The tracrRNA base pairs with the repeat regions of the pre-CRISPR RNA to form double-stranded RNA segments (Hille and Charpentier et al).

RNase III recognizes these double-stranded regions and cleaves them to produce mature crRNA. When this crRNA is connected to the enzyme Cas9, the mature crRNA and tracrRNA then associate with Cas9 to form the guide RNA complex that targets viral DNA (Hille and Charpentier et al.). One of the six domains of Cas9, the PAM Interacting Domain, allows Cas9 to identify the PAM located next to the protospacer in the viral DNA sequence, which prevents Cas9 from cutting its own CRISPR array (Anders et al.). The Histidine Asparagine Histidine domain cleaves the target strand, and the recombination and UV repair domain cleaves the non-target strand. The remaining domains, such as REC1, REC2, and Bridge Helix, contain the guide RNA (Anders et al.).

In gene editing, a single guide RNA is used instead of separate crRNA and tracrRNA. Cas9 recognizes the PAM sequence on the DNA strand, unwinds the DNA, and binds to the strand opposite to the one containing the PAM. If the protospacer sequence matches the guide RNA, Cas9 cleaves both DNA strands approximately three bases upstream of the PAM, resulting in a double-strand break. Non-homologous end joining repairs the broken DNA ends (Xue and Greene et al). Ku70 and Ku80 bind to the DNA ends, followed by DNA-dependent protein kinase catalytic subunit, XRCC4-like factor, and X-ray repair cross-complementing protein,

which help recruit DNA ligase to seal the DNA backbone. This process can introduce mutations that lead to gene knockout (Xue and Greene et al).

It also has another repair system called Homology-Driven Repair (Boutin et al.). This includes Synthesis-Dependent Strand Annealing and Double-Strand Break Repair, and it can introduce precise genomic changes, such as point mutations, when a suitable donor template is provided. Point mutations can be introduced by guiding Cas proteins to a specific genomic region with a guide RNA (Jiang and Doudna et al). The editable sites are limited because Cas proteins require a PAM sequence near the target and base editing or Homology Directed Repair can modify bases only within a specific distance from the PAM (Jiang and Doudna). Therefore, PAM availability restricts where point mutations can be generated.

Synthesis Dependent Strand Annealing uses enzymes such as the Recombination proteins B, C, and D complex, which process DNA until reaching Chi sites (Liao et al.). By providing a synthesized donor DNA with homologous sequences, new DNA can be inserted into the genome. Double Strand Break Repair can also be used to insert DNA sequences into both strands simultaneously (Liao et al.).

1.2 Other Cas system

The CRISPR-Cas system has multiple types depending on the Cas protein. Cas proteins further than Cas9 with different characteristics are developed (Table1) (Zhang). Cas12 detects PAM, which codes TTTV, uses only crRNA as a guide RNA, and the size is approximately 130 kDa, while Cas9 detects PAM coding NGG, uses crRNA and tracrRNA as guide RNA, and the size is approximately 160 kDa. Cas12 is easier to carry since it is smaller in size. In addition, it has less risk of off-target effects thanks to its PAM detection.

Cas proteins are being developed further, though they are not yet used in practical treatment. Cas12f is an exceptionally small nuclease of approximately 400 to 700 amino acids, which is approximately one-third the size of Cas9 or Cas12 (Harrington et al.; Xin et al.). Cas12f recognizes double-stranded DNA targets containing the PAM sequence TTTR (Harrington et al.). Due to its small size, it can be carried in vivo using the associated virus. However, its function is not clearly stated when it is used for eukaryotes (Xin et al.).

While Cas9 consists of about 1,400 amino acids, Casφ comprises only about 700 or 70 to 80 kDa (Pausch et al.). This compact size facilitates transport using adeno-associated virus vectors, making it highly suitable for in vivo therapies (Pausch et al.). Casφ detects the PAM sequence of TTN and has a narrow sequence, making it have less off-targeting effect.

Feature	Cas9	Cas12a	Cas12f	Casφ
Protein size	160kDa	130kDa	40 to 60kDa	70 to 80kDa

PAM recognition	5’-(A,T,C, or G)GG-3’	5’-TTT(A, C , or G)-3’	5’-TTT(A, C, or G)-3’	5’-TT(A,T ,C, or G)-3’
gRNA	crRNA + tracrRNA	crRNA	crRNA	crRNA
Off-target tendency		Moderatel y High Lower than Cas9 Very low		Presumed low (not fully characteri sed)
Editing efficiency	High in most cell types	High in moderate but depends on PAM	Low in eukaryote s	Low in mammals
Cargo size for AAV	Barely fit	Fit with small promotor	Easily fit	Easily fit
Scalability	Widely accessible	Slightly higher than Cas9	High cost	High cost (not widely available yet)
Main advantage	Well validated and efficient	Fewer off-target effect	Compact and ideal for AAV usage. Less Immune effect	Compact and AAV compatibl e

Main disadvantage	Off-target effects. Limitation in transport due to large size.	T-rich PAM limits the target.	Have narrow PAM, and have less activity	Low activity, narrow PAM, and few cell type validation s

Table1. The title of the table.

Although genome editing enables precise modification of AD-related genes, its experimental and therapeutic value depends on the availability of appropriate cellular platforms. The following section therefore, summarises the major stem cell populations used in AD research and explains how their biological properties influence downstream applications.

2. Stem Cells used in AD therapies

Stem cell-based systems have become indispensable tools for modelling AD and for exploring the feasibility of cell replacement strategies. Understanding the distinctions among MSCs, neural stem cells, and iPSCs is crucial because each population offers different advantages and constraints for both mechanistic research and therapeutic development (Table 2).

2.1 MSCs

The concept of MSC was first introduced by Alexander Friedenstein. The somatic layer of the lateral plate mesoderm gives rise to various cell types (Sheng G et al). These include the dermis, the subcutaneous layers of the body wall, cartilage, and bones in developing limbs, and a notable portion of subcutaneous fat in areas including the abdomen and thighs. This embryonic origin also produces resident progenitor cells (Sheng G et al).

These progenitor cells are often referred to as MSC-like due to their stem-like properties, and they are found in adipose and bone marrow environments. To understand the origin and regulation of MSCs, it is important to consider how the somatic mesodermal layer changes. The epithelial-to-mesenchymal transition plays a key role in guiding these cells into specialized lineages during development.

MSCs can differentiate into osteoblasts, chondrocytes, and adipocytes (C. Wang et al.). MSCs secrete various bioactive substances, including growth factors, cytokines, and extracellular vesicles (X. Han et al.). These molecules influence the surrounding microenvironment, enhance tissue regeneration, stimulate angiogenesis, and support cell survival while also reducing inflammation and promoting functional recovery.

2.2 NSCs

NSCs, originating from either embryonic or adult sources, possess the capacity to generate neurons, astrocytes, and oligodendrocytes (D. Han et al.). Through differentiation, NSCs can replenish neuronal populations lost in AD, and they also secrete neurotrophic molecules such as Brain Derived Neurotrophic Factor and Glial cell Line Derived Neurotrophic Factor, which improve synaptic plasticity, protect neurons from degeneration, and help sustain overall brain function (Pecoraro et al.; Wu et al.).

NSCs demonstrate a remarkable migratory capacity, allowing them to travel long distances and home to injured areas of the brain across all age groups (Kaminska et al.). This directed migration is mediated by the upregulation of stromal cell-derived factor 1 in astrocytes and endothelial cells within damaged tissue and by the expression of its receptor CXC chemokine receptor 4 on the surface of NSC (De Gioia et al.). However, the survival rate of NSCs after transplantation is low, with a survival rate of about thirty percent due to apoptosis caused by ischemic and hypoxic microenvironments and by immune rejection (Gowrishankar et al.).

2.3 iPSCs

iPSCs are produced from skin or fibroblast cells with the Yamanaka factors Octamer binding transcription factor 3 or 4, SRY box transcription factor 2, Kruppel-like factor 4, and Cellular myelocytomatosis oncogene added (Takahashi and Yamanaka). These iPSCs possess broad pluripotency, which enables them to give rise to multiple cell types. However, because they include genes that are activated in cancer cells, iPSCs have a small probability of becoming cancer cells due to small mutations (Liang et al.).

	NSCs	MSCs	iPSCs
Survival rate after transplantation	Moderate. Survival largely depends on the transplantation site and inflammatory microenvironment .	Relatively high. MSCs show low immunogenicity.	Low to moderate. Survival varies depending on differentiation state. Also faces immune rejection.
Degree of Differentiation control difficulty	Moderate. Differentiation is largely restricted to neural lineages. Therefore, it is relatively	Low to moderate. Differentiation into bone, cartilage, and adipose lineages is well established and easily	High. Broad pluripotency makes precise control difficult, and unwanted differentiation or teratoma

	controllable.	induced.	formation can occur.
Immunological merit and disadvantages	Possible immune rejection in allogeneic transplantation	Strong immunomodulatory capacity with low rejection risk	Autologous iPSCs have good immune compatibility, while allogeneic use may trigger immune responses.

Table2. The title.

The characteristics of these stem cell populations shape how AD can be reconstructed in vitro and how specific pathogenic mutations can be analysed. The next section examines how CRISPR-edited stem cell-derived systems have been used to model AD and to dissect its molecular mechanisms.

3. AD model using Stem Cells and CRISPR-Cas system

A major goal in AD research is the ability to recreate key pathological features in controlled experimental systems. By combining stem cell technologies with CRISPR-based editing, it has become possible to generate human cell-derived models that capture amyloid-related abnormalities, tau dysregulation, and age-associated cellular decline.

3.1 Gene Editing of the AD-causing gene

Presenilin 1 is a core component of the γ-secretase complex, which cleaves several membrane proteins, including the amyloid precursor protein. Mutations in Presenilin1 alter γ-secretase activity and lead to an increase in Aβ42, and the accumulation of Aβ in the brain will cause AD (L. Liu et al.). Presenilin2 is a gene that encodes presenilin 2, a homolog of presenilin1, and similar to presenilin1, presenilin2 forms part of the γ-secretase complex and contributes to amyloid precursor protein cleavage. Mutations in presenilin2 cause a similar but milder or later onset form of AD (L. Liu et al.).

By correcting Presenilin1 and Presenilin2 alteration, it becomes possible to avoid the excessive production of Aβ42 that causes AD. This can be achieved using the CRISPR-Cas system for gene modification. It has also been reported that CRISPR Cas9 can accurately edit the APOE4 allele. (H. Liu et al.). When the APOE4 allele in patient-derived iPSCs was converted to APOE3, the edited cells showed reduced tau phosphorylation compared to the unedited APOE4

cells, indicating that converting APOE4 to APOE3 can attenuate tau-related pathological changes (H. Liu et al.).

3.2 Artificial AD Model using stem cells

iPSC-derived NSCs have been used to establish disease models in which AD-related genes are modified using CRISPR-Cas. These NSCs can differentiate into neurons and astrocytes, making them suitable for analysing the pathological features of AD (Cao et al.). By introducing or repairing mutations in APP or PSEN1 using CRISPR Cas9, researchers have reproduced abnormalities such as altered Aβ42 production and tau pathology in great detail. In addition, the application of CRISPR-based screening to aged NSCs has allowed the identification of regulators that control the decline of NSC activity and revealed links between age-related stem cell changes and AD pathology (Ruetz et al.). In addition, iPSC-derived models enabled direct verification that APP and PSEN1 mutations lead to abnormal Aβ production and tau phosphorylation in human cells (Raffaele et al.).

3.3 Screening using CRISPR-Cas editing to make aged neural cells

The application of CRISPR-based screening to aged NSCs has allowed the identification of regulators that control the decline of NSC activity and revealed links between age-related stem cell changes and AD pathology (Ruetz et al.). CRISPR screening has also suggested that metabolic and cell cycle-related genes that are specific to aged NSCs control their activity, and some of these genes may act as drivers of neurodegeneration (Ruetz et al.).

3.4 Limitations to these models

One important advantage of this system is that it can reconstruct AD pathology using human cells. Gene editing allows direct testing of the effects of individual mutations, and CRISPR-Cas-based screens enable systematic discovery of new disease related genes (Saurat et al.). In addition, patient specific iPSCs can be used for personalised research and for potential future transplantation of gene-corrected cells. On the other hand, there are important limitations. CRISPR-Cas-based editing can cause off-target mutations, and editing efficiency can vary among experiments (Janik et al.). Another limitation is that iPSC-derived neural cells do not fully preserve age-related epigenetic changes, meaning that they cannot perfectly model the age-dependent aspects of AD pathology (Ruetz et al.). While these models have substantially improved our understanding of AD mechanisms, they also reveal barriers that must be addressed before translation to therapy is possible. The following section focuses on how stem cell and genome editing technologies are being applied in regenerative approaches and on the practical challenges that arise during transplantation and immune regulation.

4. Transplantation and risks

As research progresses from disease modelling toward therapeutic intervention, stem cell

transplantation and genome engineering approaches present both promising opportunities and significant challenges. This section evaluates current regenerative strategies, immunological barriers to cell engraftment, and the limitations of existing CRISPR platforms for AD therapy.

4.1 AD Therapy using transplantation

Transplantation studies show that neural and MSCs can lower Aβ and tau levels, boost neurogenesis and synaptic plasticity, and partially improve cognitive function. (Najm et al.)They appear to help through direct cell replacement and through the release of factors that support neurons and reduce inflammation. MSCs secrete various bioactive substances such as growth factors, cytokines, and extracellular vesicles, and these molecules influence the surrounding microenvironment, enhance tissue regeneration, stimulate angiogenesis, and support cell survival while also reducing inflammation and promoting functional recovery (Nakano et al.).

4.2 Risks

It is possible to conduct immunomodulation and reduce the risk of fibrosis when using iPSC-derived cells for transplantation. However, since iPSCs contain genes that are activated in cancer cells, they have a small probability of becoming cancer cells due to a small mutation (Liang et al.).

4.3 Prevention of immune rejection

iPSCs require the knockout of HLA genes to prevent immune rejection (Tapia and Schöler). HLA genes on chromosome 6 encode highly polymorphic molecules that present antigens on the cell surface, and these molecules enable the immune system to distinguish self from non-self (Robinson et al.). When donor cells express mismatched HLA molecules, recipients can recognize them as non-self and activate lymphocyte-mediated immune responses, which trigger immune rejection. Knockout of HLA genes, therefore, reduces the immunogenicity of iPSCs and allows them to evade allogeneic immune attack. HLA deletion can be achieved through CRISPR-Cas gene editing.

Although this strategy has not yet been applied to AD type brain tissues or organoids, hypoimmunogenic HLA edited stem cells have already been used in other transplantation models such as vascular grafts. However, hypoimmunogenic HLA edited stem cells are not able to evade Natural Killer Cells because HLA class I knockout causes missing self (Petrus-Reurer et al.). It is known that it is possible to evade Natural Killer Cell attack through increasing gene expression of HLA E and HLA G, and although this is not used for practical treatment yet, it is likely that HLA E and HLA G gene expression increase will be used in the near future (Tsuneyoshi et al.).

4.4 Limitations of each CRISPR-Cas

CRISPR-Cas systems differ in the balance between efficiency and off-target effects. High efficiency can lead to cost-effective gene mutation. On the contrary, off-target effects can lead to

carcinogenesis (Xin et al.). CRISPR-Cas9 has high efficiency for editing human neural cells; however, it has a higher off-target effect due to a less restrictive PAM sequence (S.-W. Wang et al.).The use of CRISPR Cas12f, Cas14, and CRISPR Cas Phi may enable gene editing with fewer off-target effects, improving the safety of treatments. However, both systems have limitations because Cas12f shows low activity in eukaryotic cells and Cas Phi exhibits limited efficiency in mammalian cells. These limitations reduce the overall effectiveness of gene editing using Cas12f and Cas Phi (Harrington et al.; Xin et al.). CRISPR editing can also cause off-target mutations, and editing efficiency can vary among experiments.

Conclusion

The collaboration of stem cell technologies, including NSCs, iPSCs, and MSCs, with gene editing platforms such as the CRISPR Cas system has introduced a remarkable shift in the approach to AD treatment. However, the technologies currently available remain far from curative therapies. A careful review of these methods shows that future AD treatment will not rely on a single technology but rather on the combination of multiple strategies. Transplantation of NSCs into the hippocampus has shown significant positive effects, including reductions in Aβ42 and tau protein accumulation and the promotion of neurogenesis. These advances move us closer to practical therapeutic applications, not only through direct effects but also through indirect ones. For example, organoids and cells genetically engineered using CRISPR Cas systems serve as powerful tools for identifying disease-causing genes in AD. This contributes to further discovery of AD mechanisms and provides clearer strategies for therapeutic development.

The immediate issue we face is the limited evidence demonstrating that genetically modified cells can achieve reliable immune tolerance. Although transplantation of neural cells derived from gene-corrected APP PS1 mutant iPSCs has shown beneficial effects, immune rejection and poor survival of transplanted cells remain major limitations. Therefore, more experiments are required, both in vitro and in vivo, to prevent immune reactions not only from cytotoxic T lymphocytes but also from NK cells. It is also necessary to address the risk of malignant transformation. Transplanted cells genetically engineered for immune tolerance may acquire cancer-like properties and escape elimination by NK cells and T lymphocytes. Therefore, these cells must be equipped with a suicide program that activates when they undergo oncogenic changes.

Furthermore, although CRISPR Cas technologies have advanced substantially, further improvements are required because current systems still present limitations, particularly in balancing gene editing efficiency with reduced off-target effects. CRISPR-Cas9 generally exhibits higher editing efficiency than CRISPR-Cas12f, whereas CRISPR-Cas12f shows lower off-target activity compared to CRISPR-Cas9. This highlights the need to identify or engineer Cas proteins that achieve both characteristics, low off-target effects and high editing efficiency. Using Cas proteins with these traits would allow gene editing to be performed more safely and

more cost-effectively

Ethical considerations need to be made in AD treatment research, particularly when it involves germ cell editing of AD-related genes such as APOE4. This field remains under intense debate, and China's Regulations on Human Genetic Resources Management clearly prohibit the clinical use of germ cell gene editing, while in the U.S., basic research is permitted but is subject to strict institutional scrutiny through Institutional Review Board (IRB) and Scientific Advisory Committee (SAC) review. The scientific community broadly recognizes that modifying APOE4 makes concerns about genetic enhancement and fairness in society rise even though it could theoretically lower AD risk. Consequently, the creation of interdisciplinary ethics bodies to assess potential benefits and risks, the encouragement of transparent public discussion, the reinforcement of regulatory oversight for clinical studies, and the development of robust ethical frameworks are all critical steps to guide responsible progress in this area.

Works Cited

Adli, Mazhar. "The CRISPR Tool Kit for Genome Editing and Beyond." *Nature Communications*, vol. 9, no. 1, May 2018, p. 1911. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-018-04252-2.

Anders, Carolin, et al. "Structural Basis of PAM-Dependent Target DNA Recognition by the Cas9 Endonuclease." *Nature*, vol. 513, no. 7519, Sept. 2014, pp. 569–73. *PubMed*, https://doi.org/10.1038/nature13579.

Arayici, Mehmet Emin, and Ali Kose. "Prevalence of Alzheimer's Disease and Cardiometabolic Multimorbidity in Older Adults Aged 60 and above in a Large-Scale Representative Sample in Türkiye: A Nationwide Population-Based Cross-Sectional Study." *Journal of Epidemiology and Global Health*, vol. 15, no. 1, June 2025, p. 86. *DOI.org (Crossref)*, https://doi.org/10.1007/s44197-025-00435-5.

Bahnassawy, Lamiaa, et al. "Establishment of a High-Content Imaging Assay for Tau Aggregation in hiPSC-Derived Neurons Differentiated from Two Protocols to Routinely Evaluate Compounds and Genetic Perturbations." *SLAS Discovery*, vol. 29, no. 2, Mar. 2024, p. 100137. *DOI.org (Crossref)*, https://doi.org/10.1016/j.slasd.2023.12.009.

Bermudez, Camilo, et al. "Impact of Cardiovascular Risk Factors on Plasma Biomarkers in Prediction of Alzheimer's and Cerebrovascular Neuropathology." *The Journal of Prevention of Alzheimer's Disease*, vol. 12, no. 8, Sept. 2025, p. 100224. *DOI.org (Crossref)*, https://doi.org/10.1016/j.tjpad.2025.100224.

Boutin, Laetitia, et al. "Efficient Method for Generating Point Mutations in the Vaccinia Virus Genome Using CRISPR/Cas9." *Viruses*, vol. 14, no. 7, July 2022, p. 1559. *DOI.org (Crossref)*, https://doi.org/10.3390/v14071559.

Breijyeh, Zeinab, and Rafik Karaman. "Comprehensive Review on Alzheimer's Disease: Causes and Treatment." *Molecules*, vol. 25, no. 24, Dec. 2020, p. 5789. *DOI.org (Crossref)*, https://doi.org/10.3390/molecules25245789.

Cao, Zimeng, et al. "Promoting Alzheimer's Disease Research and Therapy with Stem Cel Technology." *Stem Cell Research & Therapy*, vol. 15, no. 1, May 2024, p. 136. *DOI.org (Crossref)*, https://doi.org/10.1186/s13287-024-03737-w.

De Gioia, Roberta, et al. "Neural Stem Cell Transplantation for Neurodegenerative Diseases." *International Journal of Molecular Sciences*, vol. 21, no. 9, Apr. 2020, p. 3103. *PubMed*, https://doi.org/10.3390/ijms21093103.

Ge, Mengqian, et al. "Role of Calcium Homeostasis in Alzheimer's Disease." *Neuropsychiatric Disease and Treatment*, vol. 18, 2022, pp. 487–98. *PubMed*, https://doi.org/10.2147/NDT.S350939.

Gowrishankar, Shravan, et al. "Immunosuppression in Stem Cell Clinical Trials of Neural and Retinal Cell Types: A Systematic Review." *PLOS ONE*, edited by Nazmul Haque, vol. 19, no. 7, July 2024, p. e0304073. *DOI.org (Crossref)*,

https://doi.org/10.1371/journal.pone.0304073.
Han, Dong, et al. “Multipotent Neural Stem Cells Originating from Neuroepithelium Exist Outside the Mouse Central Nervous System.” *Nature Cell Biology*, vol. 27, no. 4, Apr. 2025, pp. 605–18. *DOI.org (Crossref)*, https://doi.org/10.1038/s41556-025-01641-w.
Han, Xia, et al. “Mesenchymal Stem Cells in Treating Human Diseases: Molecular Mechanisms and Clinical Studies.” *Signal Transduction and Targeted Therapy*, vol. 10, no. 1, Aug. 2025, p. 262. *DOI.org (Crossref)*, https://doi.org/10.1038/s41392-025-02313-9.
Harrington, Lucas B., et al. “Programmed DNA Destruction by Miniature CRISPR-Cas14 Enzymes.” *Science (New York, N.Y.)*, vol. 362, no. 6416, Nov. 2018, pp. 839–42. *PubMed*, https://doi.org/10.1126/science.aav4294.
———. “Programmed DNA Destruction by Miniature CRISPR-Cas14 Enzymes.” *Science (New York, N.Y.)*, vol. 362, no. 6416, Nov. 2018, pp. 839–42. *PubMed*, https://doi.org/10.1126/science.aav4294.
Hille, Frank, and Emmanuelle Charpentier. “CRISPR-Cas: Biology, Mechanisms and Relevance.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 371, no. 1707, Nov. 2016, p. 20150496. *PubMed*, https://doi.org/10.1098/rstb.2015.0496.
———. “CRISPR-Cas: Biology, Mechanisms and Relevance.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 371, no. 1707, Nov. 2016, p. 20150496. *PubMed*, https://doi.org/10.1098/rstb.2015.0496.
Janik, Edyta, et al. “Various Aspects of a Gene Editing System-CRISPR-Cas9.” *International Journal of Molecular Sciences*, vol. 21, no. 24, Dec. 2020, p. 9604. *PubMed*, https://doi.org/10.3390/ijms21249604.
Jiang, Fuguo, and Jennifer A. Doudna. “CRISPR-Cas9 Structures and Mechanisms.” *Annual Review of Biophysics*, vol. 46, May 2017, pp. 505–29. *PubMed*, https://doi.org/10.1146/annurev-biophys-062215-010822.
———. “CRISPR-Cas9 Structures and Mechanisms.” *Annual Review of Biophysics*, vol. 46, May 2017, pp. 505–29. *PubMed*, https://doi.org/10.1146/annurev-biophys-062215-010822.
Kaminska, Agnieszka, et al. “Interaction of Neural Stem Cells (NSCs) and Mesenchymal Stem Cells (MSCs) as a Promising Approach in Brain Study and Nerve Regeneration.” *Cells*, vol. 11, no. 9, Apr. 2022, p. 1464. *PubMed*, https://doi.org/10.3390/cells11091464.
Killelea, Tom, et al. “Cas1-Cas2 Physically and Functionally Interacts with DnaK to Modulate CRISPR Adaptation.” *Nucleic Acids Research*, vol. 51, no. 13, July 2023, pp. 6914–26. *PubMed*, https://doi.org/10.1093/nar/gkad473.
Liang, Yi, et al. “The Propensity for Tumorigenesis in Human Induced Pluripotent Stem Cells

Related with Genomic Instability." *Chinese Journal of Cancer*, vol. 32, no. 4, Apr. 2013, pp. 205–12. *PubMed*, https://doi.org/10.5732/cjc.012.10065.

———. "The Propensity for Tumorigenesis in Human Induced Pluripotent Stem Cells Is Related with Genomic Instability." *Chinese Journal of Cancer*, vol. 32, no. 4, Apr. 2013, pp. 205–12. *PubMed*, https://doi.org/10.5732/cjc.012.10065.

Liao, Hongyu, et al. "CRISPR-Cas9-Mediated Homology-Directed Repair for Precise Gene Editing." *Molecular Therapy. Nucleic Acids*, vol. 35, no. 4, Dec. 2024, p. 102344. *PubMed*, https://doi.org/10.1016/j.omtn.2024.102344.

———. "CRISPR-Cas9-Mediated Homology-Directed Repair for Precise Gene Editing." *Molecular Therapy. Nucleic Acids*, vol. 35, no. 4, Dec. 2024, p. 102344. *PubMed*, https://doi.org/10.1016/j.omtn.2024.102344.

Liu, Hang, et al. "APOE3ch Alleviates Aβ and Tau Pathology and Neurodegeneration in the Human APPNL-G-F Cerebral Organoid Model of Alzheimer's Disease." *Cell Research*, vol. 34, no. 6, Apr. 2024, pp. 451–54. *DOI.org (Crossref)*, https://doi.org/10.1038/s41422-024-00957-w.

———. "APOE3ch Alleviates Aβ and Tau Pathology and Neurodegeneration in the Human APPNL-G-F Cerebral Organoid Model of Alzheimer's Disease." *Cell Research*, vol. 34, no. 6, Apr. 2024, pp. 451–54. *DOI.org (Crossref)*, https://doi.org/10.1038/s41422-024-00957-w.

———. "APOE3ch Alleviates Aβ and Tau Pathology and Neurodegeneration in the Human APPNL-G-F Cerebral Organoid Model of Alzheimer's Disease." *Cell Research*, vol. 34, no. 6, Apr. 2024, pp. 451–54. *DOI.org (Crossref)*, https://doi.org/10.1038/s41422-024-00957-w.

Liu, Lei, et al. "The Pathogenicity of PSEN2 Variants Is Tied to Aβ Production and Homology to PSEN1." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 20, no. 12, Dec. 2024, pp. 8867–77. *PubMed*, https://doi.org/10.1002/alz.14339.

———. "The Pathogenicity of PSEN2 Variants Is Tied to Aβ Production and Homology to PSEN1." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 20, no. 12, Dec. 2024, pp. 8867–77. *PubMed*, https://doi.org/10.1002/alz.14339.

Najm, Ramsey, et al. "In Vivo Chimeric Alzheimer's Disease Modeling of Apolipoprotein E4 Toxicity in Human Neurons." *Cell Reports*, vol. 32, no. 4, July 2020, p. 107962. *DOI.org (Crossref)*, https://doi.org/10.1016/j.celrep.2020.107962.

Nakano, Masako, et al. "Bone Marrow-Derived Mesenchymal Stem Cells Improve Cognitive Impairment in an Alzheimer's Disease Model by Increasing the Expression of microRNA-146a in Hippocampus." *Scientific Reports*, vol. 10, no. 1, July 2020, p. 10772. *DOI.org (Crossref)*, https://doi.org/10.1038/s41598-020-67460-1.

Pausch, Patrick, et al. "CRISPR-CasΦ from Huge Phages Is a Hypercompact Genome Editor." *Science (New York, N.Y.)*, vol. 369, no. 6501, July 2020, pp. 333–37. *PubMed*, https://doi.org/10.1126/science.abb1400.

———. “CRISPR-CasΦ from Huge Phages Is a Hypercompact Genome Editor.” *Science*, vol. 369, no. 6501, July 2020, pp. 333–37. *DOI.org (Crossref)*, https://doi.org/10.1126/science.abb1400.

Petrus-Reurer, Sandra, et al. “Immunological Considerations and Challenges for RegenerativeCellular Therapies.” *Communications Biology*, vol. 4, no. 1, June 2021, p. 798. *DOI.org (Crossref)*, https://doi.org/10.1038/s42003-021-02237-4.

Robinson, James, et al. “IPD-IMGT/HLA Database.” *Nucleic Acids Research*, Oct. 2019, p. gkz950. *DOI.org (Crossref)*, https://doi.org/10.1093/nar/gkz950.

Ruetz, Tyson J., et al. “CRISPR–Cas9 Screens Reveal Regulators of Ageing in Neural Stem Cells.” *Nature*, vol. 634, no. 8036, Oct. 2024, pp. 1150–59. *DOI.org (Crossref)*, https://doi.org/10.1038/s41586-024-07972-2.

———. “CRISPR–Cas9 Screens Reveal Regulators of Ageing in Neural Stem Cells.” *Nature*, vol. 634, no. 8036, Oct. 2024, pp. 1150–59. *DOI.org (Crossref)*, https://doi.org/10.1038/s41586-024-07972-2.

———. “CRISPR–Cas9 Screens Reveal Regulators of Ageing in Neural Stem Cells.” *Nature*, vol. 634, no. 8036, Oct. 2024, pp. 1150–59. *DOI.org (Crossref)*, https://doi.org/10.1038/s41586-024-07972-2.

Saurat, Nathalie, et al. “Genome-Wide CRISPR Screen Identifies Neddylation as a Regulator of Neuronal Aging and AD Neurodegeneration.” *Cell Stem Cell*, vol. 31, no. 8, Aug. 2024, pp. 1162-1174.e8. *DOI.org (Crossref)*, https://doi.org/10.1016/j.stem.2024.06.001.

Takahashi, Kazutoshi, and Shinya Yamanaka. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell*, vol. 126, no. 4, Aug. 2006, pp. 663–76. *PubMed*, https://doi.org/10.1016/j.cell.2006.07.024.

Tapia, Natalia, and Hans R. Schöler. “Molecular Obstacles to Clinical Translation of iPSCs.” *Cell Stem Cell*, vol. 19, no. 3, Sept. 2016, pp. 298–309. *DOI.org (Crossref)*, https://doi.org/10.1016/j.stem.2016.06.017.

Thakor, Vishwa S., et al. “Alois Alzheimer (1864-1915): The Father of Modern Dementia Research and the Discovery of Alzheimer’s Disease.” *Cureus*, vol. 16, no. 10, Oct. 2024, p. e71731. *PubMed*, https://doi.org/10.7759/cureus.71731.

Tsuneyoshi, Norihiro, et al. “Hypoimmunogenic Human iPSCs Expressing HLA-G, PD-L1, and PD-L2 Evade Innate and Adaptive Immunity.” *Stem Cell Research &*

Therapy, vol. 15, no. 1, July 2024, p. 193. *DOI.org (Crossref)*, https://doi.org/10.1186/s13287-024-03810-4.

Wang, Cheng, et al. “Differentiation of Bone Marrow Mesenchymal Stem Cells in Osteoblasts and Adipocytes and Its Role in Treatment of Osteoporosis.” *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 22, Jan. 2016, pp. 226–33. *PubMed*,

https://doi.org/10.12659/msm.897044.

Wang, Si-Wei, et al. “Current Applications and Future Perspective of CRISPR/Cas9 Gene

Editing in Cancer." *Molecular Cancer*, vol. 21, no. 1, Feb. 2022, p. 57. *PubMed*, https://doi.org/10.1186/s12943-022-01518-8.

Xin, Changchang, et al. "Comprehensive Assessment of Miniature CRISPR-Cas12f Nucleases for Gene Disruption." *Nature Communications*, vol. 13, no. 1, Sept. 2022, p. 5623. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-022-33346-1.

———. "Comprehensive Assessment of Miniature CRISPR-Cas12f Nucleases for Gene Disruption." *Nature Communications*, vol. 13, no. 1, Sept. 2022, p. 5623. *PubMed*, https://doi.org/10.1038/s41467-022-33346-1.

———. "Comprehensive Assessment of Miniature CRISPR-Cas12f Nucleases for Gene Disruption." *Nature Communications*, vol. 13, no. 1, Sept. 2022, p. 5623. *PubMed*, https://doi.org/10.1038/s41467-022-33346-1.

Xue, Chaoyou, and Eric C. Greene. "DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing." *Trends in Genetics: TIG*, vol. 37, no. 7, July 2021, pp. 639–56. *PubMed*, https://doi.org/10.1016/j.tig.2021.02.008.

———. "DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing." *Trends in Genetics: TIG*, vol. 37, no. 7, July 2021, pp. 639–56. *PubMed*, https://doi.org/10.1016/j.tig.2021.02.008.

Zhang, F. "Development of CRISPR-Cas Systems for Genome Editing and Beyond." *Quarterly Reviews of Biophysics*, vol. 52, 2019, p. e6. *DOI.org (Crossref)*, https://doi.org/10.1017/S0033583519000052.

Bridging the Adherence Gap: The Role of MCT Supplementation and Exogenous Ketones in Pediatric Epilepsy Management By Mustafa Tornaci

Abstract

Drug-resistant epilepsy (DRE) affects approximately 30–40% of children with epilepsy. This unmet need presents a strong case for exploring and improving alternative non-pharmacological interventions once traditional treatments prove inadequate. (Borowicz-Reutt et al.; Falsaperla et al.) The Classical Ketogenic Diet (CKD) has been a staple treatment option since 1921. CKD, for a select number of patients, can reduce seizure frequency by more than 90%. (Borowicz-Reutt et al.) The CKD historically required a rigid 4:1 lipid-to-non-lipid ratio. This rigidity leads to poor long-term adherence, exacerbating dietary monotony, and creates a significant burden on caregivers. To create a more manageable approach to the CKD, this literature review synthesized data from 15 peer-reviewed studies, including randomized controlled trials, systematic reviews, and metabolic studies published between 2022 and 2026. The purpose of this review is to examine how supplemental Medium-Chain Triglyceride (MCT) and exogenous ketone salts could simplify the ketogenic protocol for children with epilepsy. Decanoic acid (C10), a type of medium chain fatty acid, has been shown to have direct antagonist actions at the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) without relying on ketogenesis. This suggests a promising new mechanism of therapeutic potential that might not require complete carbohydrate restriction. Results from this review confirm that more liberalized dietary treatments such as Modified Atkins Diet (MAD) and Low Glycemic Index Treatment (LGIT), in combination with MCT and/or exogenous ketone supplementation, achieve seizure-reduction rates comparable to the CKD while improving long-term adherence, growth outcomes, and tolerance. These data support a new paradigm of “Precision Nutrition”, in which individual supplemental strategies provide a more flexible diet, while maintaining efficacy and prioritizing long-term adherence of treatment for children with DRE.

Introduction

Millions of children suffering from DRE struggle with ketogenic therapy. The major barrier to clinical success is not the biological resistance of the disease itself, but the immense psychological, logistical and social demands of the rigid classic 4:1 ketogenic ratio.

Thirty to 40% of the pediatric epilepsy population has DRE. (Borowicz-Reutt et al., 2024; Schoeler et al., 2023). It is necessary to have non-pharmacological interventions when traditional antiseizure medications fail (Borowicz-Reutt et al., 2024; Schoeler et al., 2023). Introduced in 1921, the "Classic" Ketogenic Diet (CKD) has been a cornerstone treatment for epilepsy. The ketogenic diet can reduce seizures by over 90% and may even promote complete seizure freedom even after the diet is discontinued. (Borowicz-Reutt et al.) Due to extreme carbohydrate restriction, poor adherence and dietary monotony are commonly seen in long-term therapy with CKD. This creates a considerable burden for both the patients and their caregivers. (Desli et al., 2022; Ray et al., 2024). Newer ketogenic diets target blood ketone levels, β-hydroxybutyrate concentration and the ketone-to-glucose ratios, rather than total carbohydrate

elimination and strict 4:1 ratios. This review focuses on supplementing this more liberalized approach with (MCT)s, such as decanoic acid (C10) and caprylic acid (C8) (Akiyama et al., 2023; Heidt et al., 2023). This supplemental approach offers a more palatable, adaptable path to effective seizure control than the rigid exclusion of carbohydrates found in CKD.

This literature review aims to analyze the data to bridge the adherence gap in pediatric DRE therapy and enhance its long-term sustainability with the integration of MCT supplementation and exogenous ketone salts without sacrificing clinical efficacy.

Established diets like the Modified Atkins and Low Glycemic Index diet provide a wider range of food choices, but these still require strict dietary adherence that many families find unsustainable (El-Shafie et al., 2023). Identifying a "supplement-first" approach is vital because it addresses the adherence issue common with children trying to adopt the strict CKD with 4:1 ratio. Long-term retention of CKD drops significantly over time. Rates as low as 33% by six months due to boredom and complexity have been reported. (Chamma et al., “Beyond Epilepsy Management”) Understanding the exact mechanism of MCT and ketone salts can help clinicians develop and apply novel “Precision Nutrition” plans. This novel approach can reduce the "caregiver difficulty score" and prevent the common social isolation in pediatric patients on CKD (Panda et al., 2024).

This literature review expects to identify a gap in current research that, if filled, can provide a novel and valuable path for treating children with DRE with MCT and ketone salts. Metabolic precision with supplementation can replace the rigid 4:1 ratio and ensure long-term sustainability of ketogenic therapy for children with DRE.

Methodology

This review is based on a narrative synthesis of 15 peer-reviewed sources that focused on pediatric DRE. The selected studies included a variety of randomized controlled trials (RCTs), systematic reviews, and metabolic animal models. The selection criteria focused on recent studies (2022–2026) that aim to improve the adherence, safety, and tolerability of ketogenic interventions in pediatric drug-resistant epilepsy (DRE). Key databases, including PubMed and Google Scholar, were utilized. The review focuses on the studies that investigate the transition from rigid dietary ratios to supplemental interventions, such as Medium-Chain Triglyceride (MCT) oil and exogenous ketones. (Chamma et al., “Beyond Epilepsy Management”; Meng and Zhou)

Clinical Landscape of Ketogenic Therapy

The Classical Ketogenic Diet (CKD) uses a 3:1 or 4:1 lipid-to-non-lipid ratio and has been utilized for DRE since 1921. In a systematic review of 14 RCTs involving 1,114 children, six studies showed statistically significant seizure reduction (>50%) in KD-treated groups compared with controls. Although effective, a large proportion of families and patients are unable to maintain the dietary restrictions long term, leading to a commonly reported “adherence gap”. (Desli et al.)

One of the key obstacles to adherence to CKD is the need for a "kitchen scale" requirement, weighing every food component. This rigidity leads to "dietary boredom" and a decrease in quality of life, evidenced by a high "caregiver difficulty score" (Panda et al.) The KIWE trial, a UK randomized trial that evaluated the efficacy and tolerability of the CKD in infants with DRE, concluded that there were no statistically significant differences in these aspects between the KD treated group and that given another antiseizure medication, thus highlighting CKD is a viable therapeutic option. (Schoeler et al.).

To address these limitations, a variety of other ketogenic treatments have been developed, including the Modified Atkins Diet (MAD) and Low Glycemic Index Treatment (LGIT). Both have demonstrated success in achieving seizure freedom rates similar to the CKD but with significant improvements in patient side effects and caregiver burden. (approx. 53%) (El-Shafie et al.).

Research involving 132 children demonstrated that a 5:2 intermittent ratio, following a standard protocol for five days and a liberalized diet for two days, resulted in a 51% mean weekly seizure reduction, nearly identical to CKD (Panda et al.). This implies that efficacy does not require rigidly maintained ratios, but rather on metabolic stability through supplementation or strategic carbohydrate management.

Recent research has explored alternative fat sources and the Polyunsaturated Fatty Acid Ketogenic Diet (PUFAKD) in particular has shown success in providing similar seizure control but offering a more palatable and heart-healthy profile for long-term pediatric use (Ray et al.).

Molecular Mechanisms of Decanoic Acid (C10) and AMPA Receptor Inhibition

Traditional approaches to the ketogenic diet assume that therapeutic mechanisms lie solely in the systemic production of ketone bodies like Betahydroxybutyrate (BHB). However, recent metabolomic studies suggest that some medium-chain fatty acids such as decanoic acid (C10), are able to exert direct neuroprotective effects through independent pathways that do not rely on the production of ketosis (Akiyama et al.; Falsaperla et al.). Medium-chain triglycerides (MCTs) are rapidly hydrolyzed into C8 (caprylic acid) and C10. Both fatty acids can cross the blood-brain barrier and interact directly with neuronal targets (Sanya et al.; De Almeida Rabello Oliveira et al.).

The most significant breakthrough in understanding MCT efficacy is the identification of C10 as a non-competitive antagonist of the AMPA receptor. AMPA receptors are responsible for fast excitatory neurotransmission in the central nervous system and overactivation has been shown to cause seizure activity (Falsaperla et al.).

Electrophysiological recordings of decanoic acid show it is dose dependent and selectively inhibits current generated by AMPA receptors. C10 binds to a distinct site on the receptor, stabilizes it in a closed state and reduces the frequency of excitatory post-synaptic potentials (Falsaperla et al.; Meng and Zhou). This mechanism is completely separate from traditional antiseizure medications and provides a biological explanation for why diets

supplemented with MCT often succeed where pharmacotherapy and standard ketogenic protocols fail (Panda et al.).

The discovery of this direct mechanism has significant implications for compliance. If seizure control is driven by the concentration of C10 in the brain rather than a total lack of glucose, the diet can be substantially more liberal. (Heidt et al.).

Research indicates that with MCT oil supplementation, therapeutic levels of C10 can be maintained even in the presence of higher carbohydrate intake (Frenser et al.). For example, studies on healthy adults and pediatric subjects show that co-ingesting MCT oil with glucose does not prevent the rise of C10 and C8 in the plasma, preserving the "metabolic shield" against seizures while allowing the child to consume a more diverse and flexible diet. (De Almeida Rabello Oliveira et al.; Panda et al.).

In addition to C10, there are also exogenous ketone salts that provide additional protection from seizures. The ketone salt supplements rapidly elevate blood ketone levels, bypassing the need for the multi-day fasting or extreme fat-loading required by the classic 4:1 ratio (Borowicz-Reutt et al.; Meng and Zhou). By combining C8 and C10 MCTs' AMPA inhibition with the energy-stabilizing effects of exogenous ketones, clinicians can create a "hybrid" approach that is easier for families to manage (Chamma et al., "Beyond Epilepsy Management").

Comparative Efficacy of Liberalized Diets and MCT Supplementation

The primary concern for clinicians transitioning from the Classical Ketogenic Diet (CKD) to a more liberalized ketogenic diet version is the loss of seizure control. Data from randomized studies suggest that the integration of Medium-Chain Triglyceride (MCT) oil acts as a metabolic "safety net," that maintains the diet's efficacy. (Desli et al.; Akiyama et al.). The CKD relies on a strict 4:1 ratio to force the liver into ketogenesis, but MCT-based diets utilize the efficient and strong ketogenic properties of C8 and C10 fatty acids to induce therapeutic ketosis at much lower total fat percentages (Meng and Zhou; Akiyama et al.).

Recent randomized trials have shown that liberalized diets, such as the Modified Atkins Diet (MAD) and Low Glycemic Index Treatment (LGIT), result in seizure reduction rates comparable to the CKD when combined with MCT oil. In a 24-month study, 60% of patients on the CKD achieved seizure freedom and 46.7% of those on the more flexible MAD reached the same seizure freedom and an additional 6.7% achieved ≥50% seizure reduction (El-Shafie et al.).

This slight difference in seizure reduction is often offset by the significant improvement in growth parameters and fewer side effects in the liberalized diet groups (El-Shafie et al.; Ray et al.). Preliminary data looking at an intermittent approach, utilizing the 5:2 LGIT ratio, achieved a 51% average reduction in weekly seizure frequency. This is statistically equivalent to the efficacy of CKD in seizure management. (Panda et al.).

Randomized studies comparing the CKD with a PUFA-enriched ketogenic diet (PUFAKD) found no significant difference in seizure control at 3, 6, 9, or 12 months (Ray et al.). The PUFAKD was significantly better tolerated, with a marked reduction in side effects such as

constipation and vomiting, which are common reasons for dropping out of the CKD (Ray et al.; Desli et al.).

These results also translate to the adult population: in a single-center open-label study of adults with drug-resistant epilepsy supplementing a standard (non-ketogenic) diet with MCT oil twice daily, Rasmussen et al. reported an estimated 42% reduction in seizure rate ($p < 0.0001$). In this study only mild gastrointestinal side effects were observed. This supports the feasibility of an MCT-supplement approach without strict dietary ratios (Rasmussen et al.). The most compelling evidence for liberalization comes from metabolomic studies assessing the "carbohydrate ceiling." Evidence shows that caprylic acid (C8) supplementation can induce significant ketogenesis regardless of whether it is taken with or without a high-carbohydrate meal (Frenser et al.). Unlike long-chain fats, which require low insulin levels to be oxidized, MCTs bypass the typical regulatory steps of fat metabolism.

This allows for a "liberalized carbohydrate intake" where the child can consume glucose-containing foods without immediately "breaking" the therapeutic ketogenic state, provided the C10-mediated AMPA receptor inhibition remains active (Heidt et al.; Panda et al.). The integration of MCT oil and exogenous ketone salts provides a more stable anti-seizure effect than the strict CKD (Borowicz-Reutt et al.; Chamma et al., “Beyond Epilepsy Management”).

Clinical Compliance, Caregiver Burden, and Long-Term Sustainability

The primary limitation of CKD is not biological resistance, but the "adherence gap". The KIWE trial demonstrated that even when a diet is medically effective, families still discontinue it due to the logistical burden. (Schoeler et al.). MCT-supplemented diets directly address this by reducing the complexity of meal preparation and the need for precision weighing (Meng and Zhou).

Recent studies have utilized the "Caregiver Difficulty Score" to measure the impact of dietary rigidity on families. Research into intermittent protocols (such as the 5:2 LGIT ratio) found that allowing a child to participate in normal family meals for just two days a week significantly lowered stress levels for parents and decreased the child’s sense of social isolation (Panda et al.). By utilizing C10 supplementation to maintain a "metabolic shield" via AMPA receptor inhibition, clinicians allow for this flexibility without risking a return of seizures (Falsaperla et al.; Panda et al.).

Beyond the logistical burden, classical diets are often limited by gastrointestinal (GI) side effects. Comparing the CKD to a Polyunsaturated Fatty Acid (PUFA) approach showed that more diverse fat sources led to higher continuation rates because children experienced fewer instances of nausea and constipation (Ray et al.). Supplementation with MCT oil and exogenous ketones further improves this profile by allowing a higher intake of fiber-rich foods that are usually prohibited in a strict 4:1 ratio and improves the long-term safety profile and patient retention (Chamma et al., “Beyond Epilepsy Management”; Desli et al.).

Conclusion

This review has demonstrated that shifting from rigid, ratio-dependent ketogenic diets to a supplement-first approach centered on MCT oil and decanoic acid significantly improves compliance while maintaining efficacy. Because C10 targets the AMPA receptor directly, it permits greater carbohydrate intake, and the resulting dietary flexibility reduces caregiver burden and social isolation. By moving away from restrictive ratios and toward targeted supplementation, clinicians can offer a sustainable, effective, and humane treatment for children with drug-resistant epilepsy.

The evidence synthesized from these 15 studies converges on three findings.First, C10 produces direct anticonvulsant activity that does not depend on systemic ketosis. (Falsaperla et al.) Second, diets such as MAD and LGIT, when paired with MCT supplementation, give seizure reduction comparable to the CKD (El-Shafie et al.; Panda et al.). And third, flexible protocols reduce caregiver stress and improve quality of life for the child (Panda et al.). These findings reveal a clear gap in current ketogenic therapy for children with epilepsy. Future research should focus on optimizing the precise dosages of C10 and exogenous ketones needed to sustain AMPA-receptor inhibition.

References

Akiyama, Mari, et al. "Comprehensive Study of Metabolic Changes Induced by a Ketogenic Diet Therapy Using GC/MS- and LC/MS-Based Metabolomics." *Seizure: European Journal of Epilepsy*, vol. 107, Apr. 2023, pp. 52–59. *DOI.org (Crossref)*, https://doi.org/10.1016/j.seizure.2023.03.014.

Borowicz-Reutt, Kinga, et al. "Ketogenic Diet in the Treatment of Epilepsy." *Nutrients*, vol. 16, no. 9, Apr. 2024, p. 1258. *DOI.org (Crossref)*, https://doi.org/10.3390/nu16091258.

Chamma, Lara, et al. "Beyond Epilepsy Management: A Narrative Review of the Health Effects of Ketogenic Diets." *Nutrition*, vol. 137, Sept. 2025, p. 112804. *DOI.org (Crossref)*, https://doi.org/10.1016/j.nut.2025.112804.

———. "Beyond Epilepsy Management: A Narrative Review of the Health Effects of Ketogenic Diets." *Nutrition*, vol. 137, Sept. 2025, p. 112804. *DOI.org (Crossref)*, https://doi.org/10.1016/j.nut.2025.112804.

De Almeida Rabello Oliveira, Marcela, et al. “Effects of Short-Term and Long-Term Treatment with Medium- and Long-Chain Triglycerides Ketogenic Diet on Cortical Spreading Depression in Young Rats.” *Neuroscience Letters*, vol. 434, no. 1, Mar. 2008, pp. 66–70. *DOI.org (Crossref)*, https://doi.org/10.1016/j.neulet.2008.01.032.

Desli, Evangelia, et al. “The Efficacy and Safety of Ketogenic Diets in Drug-Resistant Epilepsy in Children and Adolescents: A Systematic Review of Randomized Controlled Trials.” *Current Nutrition Reports*, vol. 11, no. 2, June 2022, pp. 102–16. *DOI.org (Crossref)*, https://doi.org/10.1007/s13668-022-00405-4.

El-Shafie, Ali M., et al. “Impact of Two Ketogenic Diet Types in Refractory Childhood Epilepsy.” *Pediatric Research*, vol. 94, no. 6, Dec. 2023, pp. 1978–89. *DOI.org (Crossref)*, https://doi.org/10.1038/s41390-023-02554-w.

Falsaperla, Raffaele, et al. “AMPA Receptor Modulation Through Medium-Chain Triglycerides and Decanoic Acid Supports Nutritional Intervention in Pediatric Epilepsy.” *Nutrients*, vol. 17, no. 11, May 2025, p. 1805. *DOI.org (Crossref)*, https://doi.org/10.3390/nu17111805.

Frenser, Marius, et al. “Influence of Carbohydrate Intake on Caprylic Acid (C8:0)-Induced Ketogenesis—A Systematic Review and Meta-Analysis.” *Nutrients*, vol. 16, no. 15, July 2024, p. 2456. *DOI.org (Crossref)*, https://doi.org/10.3390/nu16152456.

Heidt, Christina, et al. “Beta-Hydroxybutyrate (BHB), Glucose, Insulin, Octanoate (C8), and Decanoate (C10) Responses to a Medium-Chain Triglyceride (MCT) Oil with and without Glucose: A Single-Center Study in Healthy Adults.” *Nutrients*, vol. 15, no. 5, Feb. 2023, p. 1148. *DOI.org (Crossref)*, https://doi.org/10.3390/nu15051148.

Meng, Xianghong, and Kequan Zhou. “Exploratory Dietary Approaches for Drug-Resistant Epilepsy Beyond Standard Ketogenic Diet and Fish Oil: A Systematic Review of Preliminary Clinical Evidence.” *Neurology International*, vol. 18, no. 1, Jan. 2026, p. 9. *DOI.org (Crossref)*, https://doi.org/10.3390/neurolint18010009.

Panda, Prateek Kumar, et al. “Efficacy of Daily versus Intermittent Low Glycemic Index Therapy Diet in Children with Drug-Resistant Epilepsy: A Randomized Controlled Trial.” *Epilepsy Research*, vol. 201, Mar. 2024, p. 107322. *DOI.org (Crossref)*, https://doi.org/10.1016/j.eplepsyres.2024.107322.

Rasmussen, Emmaline, et al. “Efficacy of Supplemental MCT Oil on Seizure Reduction of Adult Drug-Resistant Epilepsy – a Single-Center Open-Label Pilot Study.” *Nutritional Neuroscience*, vol. 26, no. 6, June 2023, pp. 535–39. *DOI.org (Crossref)*, https://doi.org/10.1080/1028415X.2022.2065816.

Ray, Subhasree, et al. “Efficacy and Tolerability of Classical and Polyunsaturated Fatty Acids Ketogenic Diet in Controlling Paediatric Refractory Epilepsy – A Randomized Study.” *Epilepsy Research*, vol. 204, Aug. 2024, p. 107395. *DOI.org (Crossref)*, https://doi.org/10.1016/j.eplepsyres.2024.107395.

Sanya, E. O., et al. “Antiseizure Effects of Ketogenic Diet on Seizures Induced with Pentylenetetrazole, 4-Aminopyridine and Strychnine in Wistar Rats.” *Niger. J. Physiol. Sci.*, 2016.

Schoeler, Natasha E., et al. “Classic Ketogenic Diet versus Further Antiseizure Medicine in Infants with Drug-Resistant Epilepsy (KIWE): A UK, Multicentre, Open-Label,

Randomised Clinical Trial." *The Lancet Neurology*, vol. 22, no. 12, Dec. 2023, pp. 1113–24. *DOI.org (Crossref)*, https://doi.org/10.1016/S1474-4422(23)00370-8.

Quantitative EEG Analysis of Brain Activity Before, During, and After Nadi Shodhana Pranayama

By Maya Kumari

1. Abstract

Nadi Shodhana Pranayama, or alternate nostril breathing, represents a fundamental yogic technique utilized to modulate the autonomic nervous system and enhance cognitive focus. While existing literature documents neurophysiological shifts during regulated respiration, research often relies on long-term interventions or lacks the integration of regional cortical comparisons and robust effect-size metrics. This investigation systematically evaluates the impact of a singular Nadi Shodhana session on brain-wave oscillations by analyzing Delta (0.5-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz), and Gamma (30-45 Hz) spectral power. Data were collected from twelve participants (ten analyzed after signal-quality exclusions) using a wearable Muse-S EEG headband across three conditions: Before (baseline), During, and After practice. Signal processing using a custom Python-based Welch PSD pipeline revealed measurable group-level changes in EEG spectral activity, with statistically significant differences in selected frequency bands: an increase in slower-wave activity (Theta and Delta bands) and a reduction in high-frequency arousal (Beta and Gamma bands). Notably, relative Alpha power transitioned from a 23.2% baseline to 25.9% post-practice (*Cohen's d* = 0.82), while Beta and Gamma exhibited large negative effect sizes during the intervention. Despite individual variability, regional analyses confirmed a consistent neurophysiological shift toward a relaxed and attentive state, establishing a reproducible framework for quantifying the effects of breathing-based interventions using wearable EEG technology.

2. Background Information

Pranayama, the yogic regulation of breath, is one of the eight limbs of classical yoga and is described in foundational Hatha texts such as the *Hatha Yoga Pradipika* (Muktibodhananda, Swami) and the *Yoga Chudamani Upanishad* (Satyadharma, Swami). Within this tradition, Nadi Shodhana - literally "channel purification" - is a slow, alternate-nostril breathing technique intended to balance the Ida and Pingala energy channels and steady the mind (Satyananda, Swami). In his classical text, *Prana and Pranayama*, Swami Niranjanananda explains that regular, systematic practice gradually removes energetic blockages, steadies the breath and mind (thereby stabilizing the nervous system), and allows prana, the life force, to flow more freely. As this happens, the breath naturally becomes subtler, mental fluctuations decrease, concentration improves, and the practitioner becomes prepared for meditation and higher pranayama. Modern physiological research frames the same practice in terms of the autonomic nervous system: slow, rhythmic breathing shifts autonomic balance toward parasympathetic (calming) activity and has been associated with reduced blood pressure, lower anxiety, and improved attention and vigilance performance (Telles and Singh).

Electroencephalography (EEG) is well suited to studying these effects because it measures synchronized cortical electrical activity with millisecond temporal resolution, capturing the rapid shifts in arousal and relaxation that breathing practices produce (Chaddad). Prior work on meditation and breath control indicates that relaxed states increase slower oscillations, particularly alpha (8–12 Hz) and theta (4–8 Hz), while arousing or effortful states enhance beta (12–30 Hz) and gamma (30–45 Hz) activity *(Cahn and Polich, 2006)*. Nadi Shodhana is a particularly informative candidate for EEG study because, unlike stimulating practices such as Bhastrika or Kapalbhati, it is explicitly a *balancing* technique - making it a natural test case for a before/during/after contrast between relaxation and arousal.

Existing EEG findings on nostril breathing are mixed, largely because of differences in how the breathing was performed. Stancák and Kuna *(International Journal of Psychophysiology, 1994)* reported that ten minutes of forced (fast) alternate-nostril breathing increased beta and partly alpha power and reduced beta hemispheric asymmetry, suggesting a balancing but mildly arousing effect. In contrast, studies of slow yogic alternate-nostril breathing report calming shifts: Telles found decreased theta during the practice and lower beta afterward, with no change in hemispheric symmetry *(BMC Research Notes, 2017)*, and Niazi *(Scientific Reports, 2022)* found that left-nostril breathing produced a posterior, relaxation-associated rhythm linked to parasympathetic activity.

Over a longer time course, Verma *(Indian Journal of Positive Psychology, 2022)* reported a significant increase in alpha EEG after 24 days of regular Nadi Shodhan practice. Taken together, forced and fast breathing tends to raise beta, whereas slow practices lower it and elevate alpha - the pattern the present study set out to test acutely, in a single session, using a wearable EEG device. In the present study, Delta and Theta increased during practice (+2.44 and +1.45 points) while Beta and Gamma decreased (-2.26 and -2.16 points), and Alpha remained 2.62 points above baseline afterward - a relaxation-dominant pattern.

3. **Scientific Question**

How does Nadi Shodhana Pranayama (Alternate Nostril Breathing) influence EEG brain-wave activity before, during, and after practice?

4. **Hypothesis**

If participants perform Nadi Shodhana Pranayama, Alpha (8-12 Hz) and Theta (4-8 Hz) power will increase while Beta (12-30 Hz) power will decrease relative to baseline, reflecting a shift toward a more relaxed and attentive brain state.

5. **Methods**

5.1 **Subjects and Recruitment**

Twelve healthy adult participants were recruited from attendees and practitioners at the *Bihar School of Yoga, Munger, India*, and *Satyananda Yoga Farm, Gilroy, USA*. All participants were healthy adults with no history of neurological or respiratory conditions. Before participation, the study procedures were explained, and written informed consent was obtained from each participant. All participants' data were anonymized prior to analysis. Of the twelve recordings, two were excluded from analysis due to missing values (NaNs) in the EEG signal, leaving ten participants for the group-level analysis.

Subject ID	Gender (M/F)	Age	Duration (Seconds)
BSY01	M	49	900
BSY02	M	55	900
BSY03	F	53	900
BSY04	M	35	900
BSY05	F	47	900
BSY06	F	42	900
BSY07	F	45	900
BSY08	F	45	900
BSY09	F	46	900
BSY10	M	52	900
SYF01	M	52	900
SYF02	M	54	900

Table 1: Participant demographics, including subject ID, gender, age, and total recording duration (seconds; 3 sessions, 15 minutes each). Twelve participants were recruited (ten from the Bihar School of Yoga, Munger, and two from Satyananda Yoga Farm, Gilroy); two recordings were excluded from statistical analysis due to missing values (NaNs) in the EEG signal, leaving ten participants for the group analysis.

5.2 **Equipment**

EEG was recorded using the Muse-S headband *(InteraXon Inc.),* a non-invasive, dry-electrode wearable device. The Muse-S records from four EEG channels positioned at AF7 and AF8 (frontal region) and TP9 and TP10 (temporal region, behind the ears), with a reference electrode at Fpz, following the international 10–20 system. Signals were sampled at 256 Hz and streamed to Mind Monitor, which was configured to capture raw EEG data.

5.3 **Pranayama Protocol**

Participants were seated comfortably, and the Muse-S headband was fitted and its signal quality checked. The Nadi Shodhana protocol was followed *(Saraswati, Asana Pranayama Mudra Bandha).* One breathing cycle consisted of inhalation through the left nostril, exhalation through the right nostril, inhalation through the right nostril, and exhalation through the left nostril, using a 1:1 inhalation-to-exhalation ratio. The technique was communicated to participants verbally by an instructor. Each session consisted of three phases: a 3-minute Before baseline of relaxed natural breathing, 5-10 minutes During Nadi Shodhana, and a 3-minute After recovery of relaxed natural breathing.

5.4 **Analysis Methods**

Raw EEG recordings were exported as CSV files and processed using a custom Python pipeline. Preprocessing included trimming the first and last 3 seconds of each recording to remove onset and offset artifacts, applying a 50 Hz notch filter to remove line noise, and band-pass filtering the signal between 1 and 45 Hz. Artifacts were further handled by removing segments flagged by the Muse signal-quality indicator. Power spectral density was then estimated using Welch's method with a 3-second window and 50% overlap. Absolute and relative band power were computed for five frequency bands: Delta (0.5-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz), and Gamma (30-45 Hz), with relative power expressed as each band's percentage of total power. Participant-level averages were computed for each condition, and regional comparisons were performed for Front vs. Back and Left vs. Right electrode pairs.

5.5 **Statistics**

For each frequency band and condition comparison (Before vs. During, During vs. After, and Before vs. After), normality was first assessed using the Shapiro-Wilk test. When the normality assumption was met (Shapiro-Wilk $p \geq 0.05$), a paired t-test was applied; otherwise, the Wilcoxon signed-rank test was used. Group-level results are reported as the mean relative power with standard deviation (SD), standard error of the mean (SEM), and 95% confidence interval (CI). The magnitude of each change was quantified using Cohen's d

effect size. Statistical significance was defined as $p < 0.05$.

6. Results

EEG analyses were performed at both the individual and group levels to evaluate changes in brain-wave activity associated with Nadi Shodhana Pranayama. Results are presented using participant summaries, group averages, regional comparisons, statistical analyses, and effect-size measurements.

6.1 Individual Results

To illustrate the range of individual responses underlying the group averages, two representative participants are presented as dashboards (Figures 1 and 2). Each dashboard reports average band power per condition (panels A, E), overall change from baseline (B, F), Front vs. Back regional differences (C, G), and Left vs. Right hemispheric differences (D, H), giving a full view of each participant's response before, during, and after Nadi Shodhana Pranayama. Both subjects showed the core group pattern during practice: Theta increased (+1.4 and +1.8 points; panels 1B and 2F) while Beta and Gamma decreased (Beta -2.0 and -1.0 points; Gamma -0.9 and -1.4 points). The two subjects differed most in Alpha - Subject 1's exhibited a decrease in Alpha during and after practice (-3.2 and -3.1 points; Figure 1B), whereas Subject 2's increased after practice (+2.6 points; Figure 2F) - illustrating that although the group-level direction held for most bands, the magnitude, and sometimes the direction, of individual responses varied.

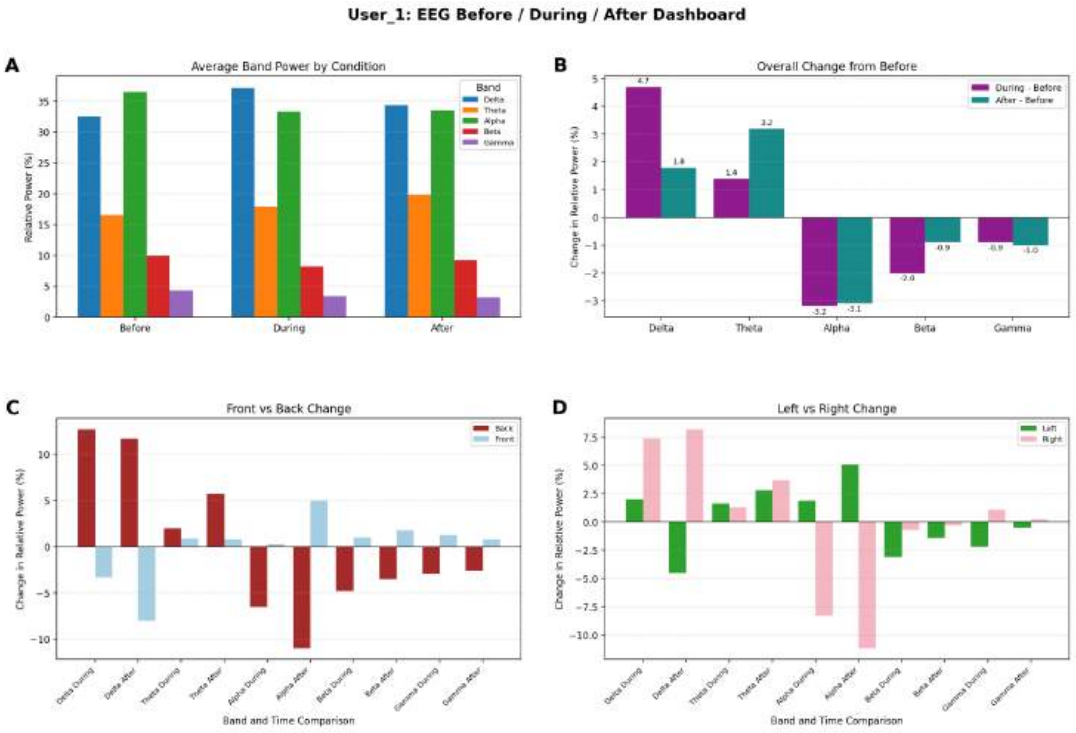


Figure 1: EEG dashboard for Subject 1, before, during, and after Nadi Shodhana Pranayama. **(A)** Average band power per condition - before, during, and after pranayama - with relative power (%) shown across delta (blue), theta (orange), alpha (green), beta (red), and gamma (purple). **(B)** Overall change in relative power from baseline for each band (During - Before, purple; After - Before, teal); values are labeled above each bar. **(C)** Front vs. Back regional change, showing back (dark red) versus front (light blue) channels for each band and time point. **(D)** Left vs. Right hemispheric change, showing left (green) versus right (pink) channels. Notably, panel B shows that Delta and Theta increased during practice (+4.7 and +1.4 points) while Beta and Gamma decreased (-2.0 and -0.9 points); unlike the group trend, this participant's Alpha *decreased* during and after practice (-3.2 and -3.1 points), illustrating the individual variability underlying the group averages.

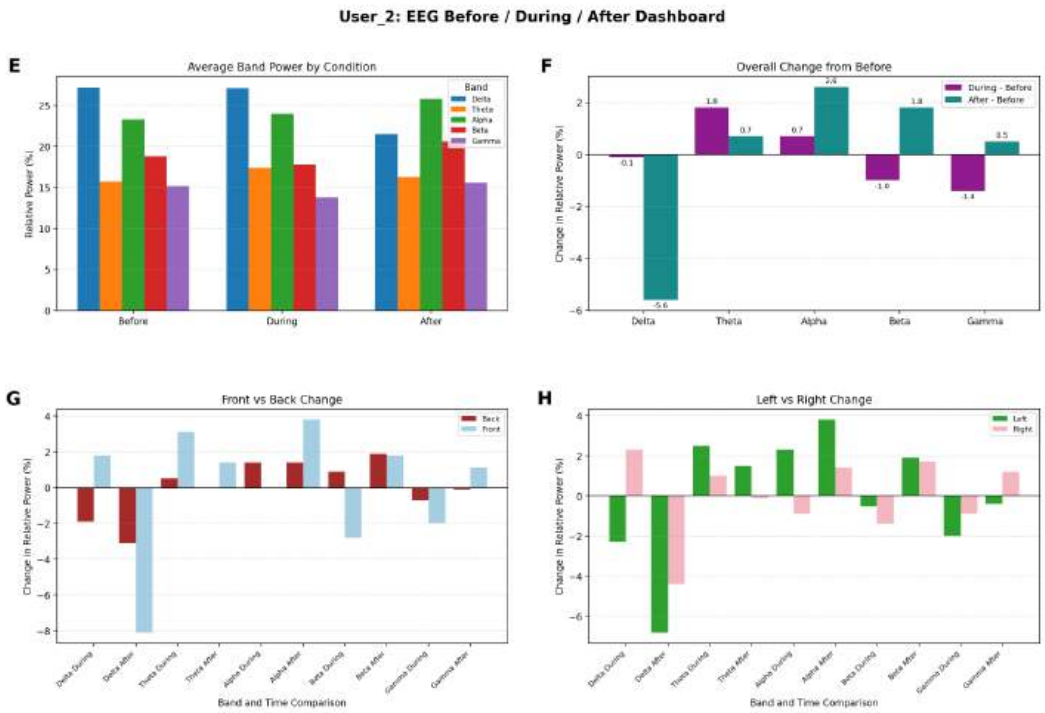


Figure 2: EEG dashboard for Subject 2, before, during, and after Nadi Shodhana Pranayama. **(E)** Average band power per condition - before, during, and after pranayama - with relative power (%) shown across delta (blue), theta (orange), alpha (green), beta (red), and gamma (purple). **(F)** Overall change in relative power from baseline for each band (During - Before, purple; After - Before, teal); values are labeled above each bar. **(G)** Front vs. Back regional change, showing back (dark red) versus front (light blue) channels for each band and time point. **(H)** Left vs. Right hemispheric change, showing left (green) versus right (pink) channels. Notably, panel F shows that Theta increased during practice (+1.8 points) while Alpha rose after practice (+2.6 points), and Beta and Gamma decreased during practice (-1.0 and -1.4 points) - the same relaxation-dominant pattern seen at the group level.

6.2 **Combined Summary Across All Subjects**

To evaluate the overall effects of Nadi Shodhana Pranayama, EEG results from all participants were combined to obtain relative band-power changes across the Before, During, and After conditions. Group-level analysis showed that during the breathing practice, relative Delta and Theta power increased by 2.44 and 1.45 points, respectively (Theta $p = 0.010$), while Beta and Gamma decreased by 2.26 and 2.16 points ($p = 0.005$ and $p = 0.013$). Following practice, Alpha power remained 2.62 points above baseline ($p = 0.030$), indicating a sustained relaxation response. These changes were observed despite individual variability and are summarized in Figure 3.

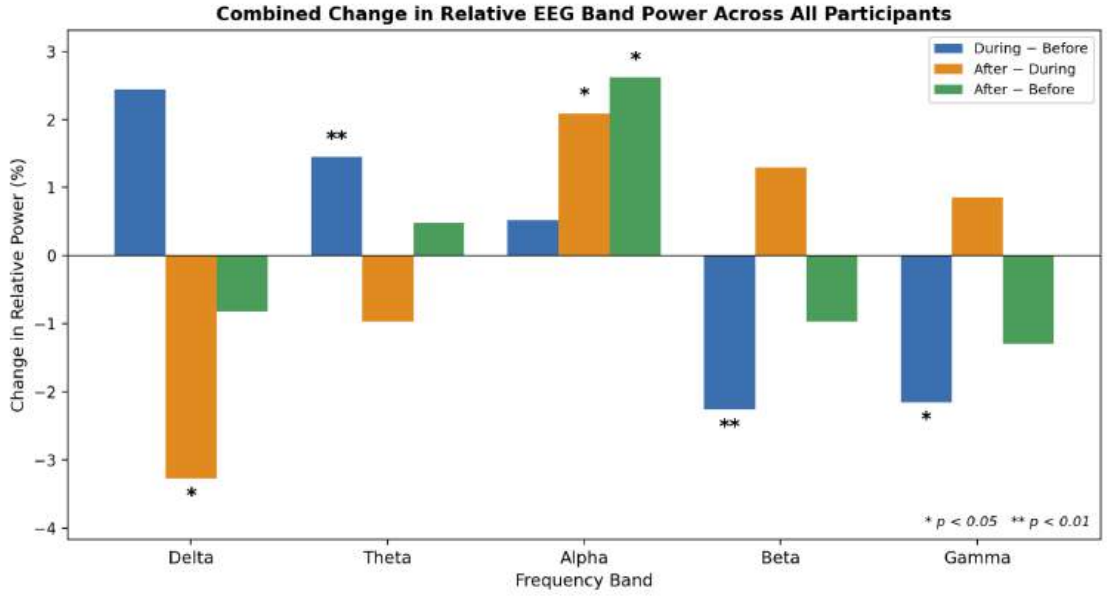


Figure 3: Combined change in relative EEG band power across all participants for three condition contrasts - During - Before (blue), After - During (orange), and After - Before (green) - for each frequency band. Positive values indicate an increase in relative power; the y-axis is the change in relative power (percentage points). The asterisks mean that statistically significant changes (* $p < 0.05$, ** $p < 0.01$) from the paired tests in Table 3. Theta increased and Beta and Gamma decreased significantly during practice (During - Before), while Alpha was significantly higher after practice relative to both the During and Before conditions.

6.3 **Group Statistics Across All 10 Participants**

The group mean relative EEG power together with the Standard Deviation (SD), Standard Error of the Mean (SEM), and 95% Confidence Interval (CI) for each frequency band across the Before, During, and After conditions. This can be viewed in Table 2.

Group Relative Power Summary: Mean, SD, SEM, and 95% CI

	Band	Condition	N	Mean_Rel%	SD	SEM	95% CI
0	Delta	Before	10	28.865	2.550	0.807	1.581
1	Delta	During	10	31.309	6.305	1.994	3.908
2	Delta	After	10	28.036	5.460	1.727	3.384
3	Theta	Before	10	16.819	2.212	0.699	1.371
4	Theta	During	10	18.266	3.363	1.063	2.084
5	Theta	After	10	17.297	2.606	0.824	1.616
6	Alpha	Before	10	23.242	5.193	1.642	3.219
7	Alpha	During	10	23.767	4.660	1.474	2.889
8	Alpha	After	10	25.857	5.283	1.670	3.274
9	Beta	Before	10	17.617	2.993	0.947	1.855
10	Beta	During	10	15.356	3.680	1.164	2.281
11	Beta	After	10	16.652	4.103	1.297	2.543
12	Gamma	Before	10	13.457	3.874	1.225	2.401
13	Gamma	During	10	11.301	4.459	1.410	2.764
14	Gamma	After	10	12.157	4.822	1.525	2.989

Table 2: Group statistics for relative EEG band power across all 10 participants before, during, and after Nadi Shodhana Pranayama.

6.4 **Regional Brain Analysis Across All 10 Participants**

The Front-Back and Left-Right comparisons showed the same directional changes across regions and hemispheres (Figure 4). During practice, Delta and Theta increased and Beta and Gamma decreased in both frontal and posterior channels and in both hemispheres, and after practice Alpha increased across all regions. The changes were consistent in direction everywhere but generally larger frontally and on the left - for example, during practice Delta rose by 3.24 points frontally versus 1.48 posteriorly, and Beta fell by 2.92 points frontally versus 1.48 posteriorly, while after practice Alpha increased by 3.35 points frontally versus 1.67 posteriorly. This indicates that the relaxation-dominant shift was not localized to one region or hemisphere but occurred across the cortex, with a frontal and left-sided emphasis.

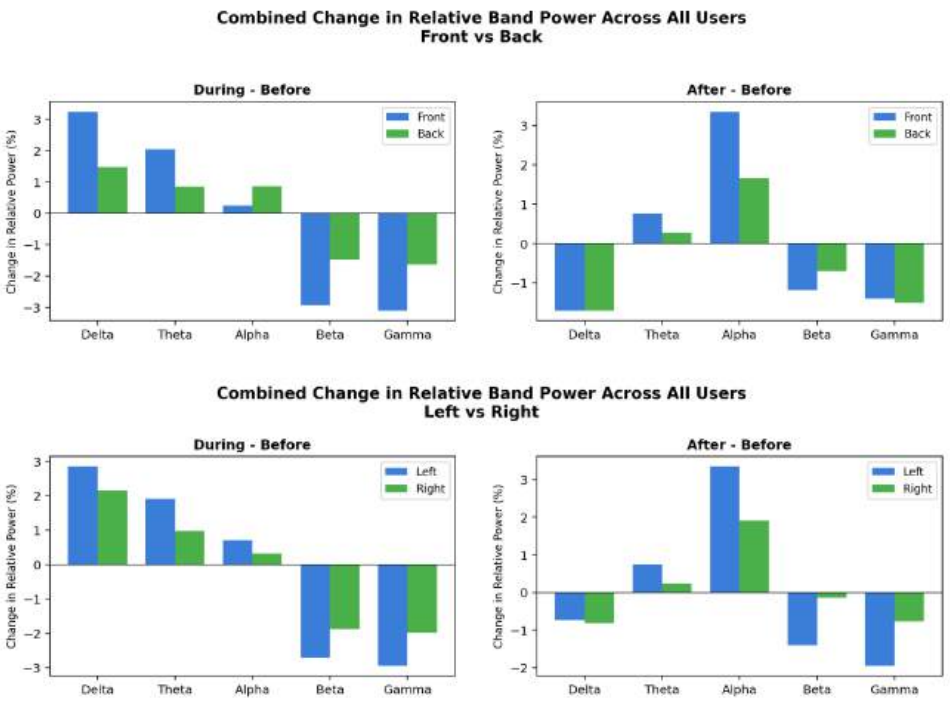


Figure 4: Regional changes in relative EEG band power across the Front–Back and Left-Right brain regions during and after Nadi Shodhana Pranayama.

6.5 **Statistical Testing Across All Participants**

Paired comparisons identified several statistically significant changes across conditions (Table 3). During practice, relative Theta power increased significantly by 1.45 points relative to baseline (Wilcoxon $p = 0.010$), while Beta and Gamma decreased significantly by 2.26 points ($p = 0.005$) and 2.16 points ($p = 0.013$), respectively. The Delta increase during practice (+2.44 points) did not reach significance ($p = 0.14$). In the recovery period, Alpha power was significantly higher than both baseline (+2.62 points; $p = 0.030$) and the During condition (+2.09 points; $p = 0.023$), and Delta decreased significantly from During to After (-3.27 points; $p = 0.030$). All remaining comparisons were non-significant ($p > 0.05$).

Paired Statistical Tests for Relative EEG Band Power

	Band	Comparison	N	Mean Change	Shapiro p	Test	Statistic	p-value	Significance
0	Delta	Before vs During	10	2.4443	0.9176	Paired t-test	1.6132	0.1412	Not significant
1	Delta	During vs After	10	-3.2729	0.8683	Paired t-test	-2.5755	0.0299	* p < .05
2	Delta	Before vs After	10	-0.8285	0.0914	Paired t-test	-0.5481	0.5970	Not significant
3	Theta	Before vs During	10	1.4464	0.0042	Wilcoxon signed-rank	3.0000	0.0098	** p < .01
4	Theta	During vs After	10	-0.9682	0.7707	Paired t-test	-1.5525	0.1550	Not significant
5	Theta	Before vs After	10	0.4783	0.8754	Paired t-test	1.0199	0.3344	Not significant
6	Alpha	Before vs During	10	0.5252	0.9171	Paired t-test	0.7563	0.4688	Not significant
7	Alpha	During vs After	10	2.0901	0.6553	Paired t-test	2.7482	0.0225	* p < .05
8	Alpha	Before vs After	10	2.6153	0.5443	Paired t-test	2.5820	0.0296	* p < .05
9	Beta	Before vs During	10	-2.2606	0.1940	Paired t-test	-3.6354	0.0054	** p < .01
10	Beta	During vs After	10	1.2957	0.1635	Paired t-test	2.1542	0.0596	Not significant
11	Beta	Before vs After	10	-0.9649	0.1247	Paired t-test	-1.0878	0.3050	Not significant
12	Gamma	Before vs During	10	-2.1554	0.8692	Paired t-test	-3.0688	0.0134	* p < .05
13	Gamma	During vs After	10	0.8552	0.6465	Paired t-test	1.5852	0.1474	Not significant
14	Gamma	Before vs After	10	-1.3002	0.9403	Paired t-test	-1.7032	0.1227	Not significant

Table 3: Paired statistical tests comparing relative EEG band power across the Before vs. During, During vs. After, and Before vs. After comparisons. The Wilcoxon signed-rank test was used when the normality assumption was not met (Shapiro-Wilk $p < 0.05$); otherwise, a paired t-test was applied. Significance: * $p < 0.05$, ** $p < 0.01$.

6.6 **Effect Size Analysis**

Effect size measures the magnitude of EEG changes, providing information about the practical importance of the observed differences. Cohen's d effect was calculated for each comparison (**Before vs. During** and **During vs. After**). This can be seen in Table 4.

Effect Size Summary

Band	Comparison	Cohen's d	Effect Size
Delta	Before vs During	0.510	Moderate
Delta	During vs After	-0.814	Large
Delta	Before vs After	-0.173	Negligible
Theta	Before vs During	0.837	Large
Theta	During vs After	-0.491	Small
Theta	Before vs After	0.323	Small
Alpha	Before vs During	0.239	Small
Alpha	During vs After	0.869	Large
Alpha	Before vs After	0.816	Large
Beta	Before vs During	-1.150	Large
Beta	During vs After	0.681	Moderate
Beta	Before vs After	-0.344	Small
Gamma	Before vs During	-0.970	Large
Gamma	During vs After	0.501	Moderate
Gamma	Before vs After	-0.539	Moderate

Table 4: Cohen's d effect sizes for changes in relative EEG band power across the **Before vs. During** and **During vs. After** conditions

7. Discussion

7.1 Interpretation on Group Findings

Effect-size and statistical analyses demonstrated meaningful EEG changes associated with Nadi Shodhana Pranayama. During the breathing practice, relative Theta power increased by 1.45 percentage points (16.82% to 18.27%; Wilcoxon $p = 0.010$, $d = 0.84$), while Beta decreased by 2.26 points (17.62% to 15.36%; $p = 0.005$, $d = -1.15$) and Gamma decreased by 2.16 points (13.46% to 11.30%; $p = 0.013$, $d = -0.97$), indicating a shift toward slower brain-wave activity and reduced cortical arousal. Delta also rose by 2.44 points during practice (28.87% to 31.31%), though this change was not statistically significant ($p = 0.14$, $d = 0.51$). After practice, Alpha showed a large, sustained increase of 2.62 points above baseline (23.24% to 25.86%; $p = 0.030$, $d = 0.82$) and remained elevated relative to the During condition (+2.09 points; $p = 0.023$, $d = 0.87$), suggesting continued relaxed wakefulness. Together, these findings indicate that Nadi Shodhana produces measurable neurophysiological changes consistent with a relaxed and attentive brain state.

8. Conclusion

The findings of this study demonstrate that Nadi Shodhana Pranayama produces measurable changes in EEG brain-wave activity. Across the group, the practice was associated with increased slow-wave activity during the intervention - Delta rose by 2.44 points and Theta by 1.45 points ($p = 0.010$) - alongside significant decreases in Beta (-2.26 points; $p = 0.005$) and Gamma (-2.16 points; $p = 0.013$). Following practice, Alpha power increased by 2.62 points above baseline ($p = 0.030$) and remained elevated, indicating a sustained relaxation response. These findings are consistent with previous research suggesting that controlled yogic breathing influences autonomic nervous system regulation and cortical activity. Although individual responses varied in magnitude, the overall trends remained consistent across participants and regional brain analyses. This study also presents a reproducible EEG analysis pipeline using wearable technology, demonstrating that Nadi Shodhana can be objectively evaluated through quantitative EEG measures and providing a foundation for future research on breathing-based interventions for relaxation, attention, and brain health.

9. Limitations

This study included 10 participants and used a four-channel wearable EEG system. Although artifact handling procedures were applied, wearable EEG devices remain susceptible to movement and electrode-contact noise. The small sample size is also a limitation. With only ten participants, the test used to check whether the data was normally distributed (Shapiro-Wilk) was not very reliable, so the choice of statistical test should be viewed with some caution. Many comparisons were also run at once without adjusting for this, which raises the chance of a result appearing significant by chance. For these reasons, the significant findings should be seen as

early and in need of confirmation. Future studies involving larger participant populations, standardized breathing durations, and additional physiological measurements such as heart rate variability or functional near-infrared spectroscopy may provide a more comprehensive understanding of the neurophysiological effects of Nadi Shodhana Pranayama.

Works Cited

Niranjanananda, Swami. *Prana and Pranayama*. Yoga Publications Trust, 2018.

Satyananda, Swami. *Asana Pranayama Mudra Bandha*. Yoga Publications Trust, 2017.

Satyananda, Swami. *A Systematic Course in the Ancient Tantric Techniques of Yoga and Kriya*. Yoga Publications Trust, 2017.

Bihar School of Yoga. "Effects of Pranayama on the Brain and Nervous System." *Yoga Magazine*, www.yogamag.net. Accessed 19 Oct. 2025.

Satyadharma, Swami. *Yoga Chudamani Upanishad*. Yoga Publications Trust, 2003.

Muktibodhananda, Swami. *Hatha Yoga Pradipika*. Yoga Publications Trust, 2018.

Cahn, B. Rael, and John Polich. "Meditation States and Traits: EEG, ERP, and Neuroimaging Studies." *Psychological Bulletin*, vol. 132, no. 2, 2006, pp. 180–211.

Chaddad, Ahmad, et al. "Electroencephalography Signal Processing: A Comprehensive Review and Analysis of Methods and Techniques." *Sensors*, vol. 23, no. 14, 2023, article 6434, https://pubmed.ncbi.nlm.nih.gov/37514728/. Accessed 12 July 2026.

InteraXon Inc. "Muse S." *Muse*, choosemuse.com/products/muse-s. Accessed 17 July 2026.

Niazi, Imran Khan, et al. "EEG Signatures Change during Unilateral Yogi Nasal Breathing." *Scientific Reports*, vol. 12, no. 1, 2022, article 520, www.nature.com/articles/s41598-021-04461-8. Accessed 18 July 2026

Stancák, Andrej, and Miloslav Kuna. "EEG Changes during Forced Alternate Nostril Breathing." *International Journal of Psychophysiology*, vol. 18, no. 1, 1994, pp. 75–79, www.sciencedirect.com/science/article/abs/pii/0167876084900175. Accessed 18 July 2026

Telles, Shirley, et al. "Hemisphere Specific EEG Related to Alternate Nostril Yoga Breathing." *BMC Research Notes*, vol. 10, no. 1, 2017, article 306, link.springer.com/article/10.1186/s13104-017-2625-6.

Telles, Shirley, and Nilkamal Singh. *Science of Yoga: Understanding Its Biological and Psychological Effects*. Swami Vivekananda Yoga University Press, 2018.

Verma, Avichal, and Sandeep Singh. "The Impact of Nadi Shodhan and Pranakarshan Pranayama on α-EEG among College Going Students." *Indian Journal of Positive Psychology*, vol. 13, no. 2, 2022, pp. 121–125, journals.indexcopernicus.com/api/file/viewByFileId/1588829. Accessed 18 July 2026.

How Can Poison Become Medicine? Medicine and Pharmaceutical Processing in Medieval China through the Rise and Fall of Five-Stone Powder (五石散) By Wushi San

Abstract

Five-Stone Powder was a dangerous medicine made from toxic minerals. In medieval China, however, it was regarded as both a powerful treatment and a health-enhancing elixir. It was most popular and widely used during the Six Dynasties period (220–589 CE), but it seems to have suddenly disappeared from historical records during the first half of the Song dynasty (960–1279). This article argues that its decline was not simply the result of prohibition, medical failure, or its excessively high cost. Rather, it formed part of a broader transition from medicine focused on managing side effects after patients had taken a drug to medicine centered on selecting, processing, and standardizing medicinal ingredients before use in order to reduce harmful toxicity while preserving their therapeutic properties.

Research Questions and Methods

This article addresses two main questions. First, why did people continue to use Five-Stone Powder for such a long time even though its serious side effects were widely known? Second, did it disappear after the Song dynasty simply because people discovered that it was toxic?

To answer these questions, this study compares medical writings and historical records produced over approximately one thousand years, from the third to the twelfth century. Its major sources include the *Essential Prescriptions of the Golden Coffer* (*Jingui Yaolüe,* 金匱要略), which records medical formulas; A *New Account of the Tales of the World* (*Shishuo Xinyu*, 世說新語) and the *History of the Southern Dynasties* (*Nan Shi*, 南史), which preserve stories about historical figures; the *Treatise on the Origins and Symptoms of Diseases* (*Zhubing Yuanhou Lun*, 諸病源候論), which analyzes the causes of illness; the *Essential Prescriptions Worth a Thousand in Gold* (*Qianjin Yaofang*, 千金要方) and the *Formulary of the Pharmacy Service for Benefiting the People in the Taiping Era* (*Taiping Huimin Hejiju Fang*, 太平惠民和齊局方), which discuss the selection and processing of medicinal ingredients; and the *Brush Talks from Dream Brook* (*Mengxi Bitan*, 夢溪筆談), which critically examines ancient prescriptions.

It is also important to recognize that these sources have different characteristics. Medical texts show actual prescriptions and medical theories, but we cannot assume that every recorded formula was used in exactly the same way throughout China. Dramatic healing stories in historical works reveal contemporary ideas, but they may have been exaggerated to make certain individuals appear more remarkable. When a politician described his own illness, he may also have been trying to avoid an official appointment. Therefore, this article does not treat any single source as complete proof. Instead, it focuses on broader changes that appear across several different sources.

Another important point is that traditional medical concepts such as “heat (熱),” “kidneys (腎),” and “qi (氣)” should not be directly equated with modern ideas of body temperature, internal organs, or energy. Physicians in the past observed and explained symptoms within their own systems of knowledge. This article does not claim that their

explanations agree with modern science. Its purpose is to understand historically how people recognized medical risks and what observations and techniques they developed to reduce treatment failures.

The history of Five-Stone Powder can be divided into three general stages. In the first stage, the main concern was managing the patient's daily life so that the heat produced by the medicine could escape from the body. In the second stage, physicians began to analyze the toxicity of the medicine itself and the different reactions of individual bodies. In the third stage, they attempted to control the properties of medicinal ingredients before use by selecting, washing, grinding, and heating them. They also began to turn these methods into shared standards. In short, these stages did not have clear boundaries, and they often overlapped. Nevertheless, they show how the main focus of medical attention gradually shifted.

Introduction: Why Were Dangerous Minerals Used as Medicine?

"Everything is poison, and nothing is without poison. The dose alone makes a thing not a poison." This statement, commonly attributed to the sixteenth-century physician Paracelsus, suggests that the boundary between medicine and poison is not as simple as it may appear. The same substance can become either a treatment or a dangerous poison depending on how much is used and how it is processed. Even minerals such as mercury, whose toxicity is now well known, were once important medicinal ingredients.

Minerals had long been used in East Asian medicine. Ancient medical texts divided medicinal substances into three categories: upper-grade, middle-grade, and lower-grade drugs. Upper-grade drugs (上藥) were believed to nourish the body and extend life. Middle-grade drugs(中藥) were used to regulate the body's balance, while lower-grade drugs(下藥) treated particular diseases. Interestingly, toxic minerals such as cinnabar (mercury sulfide, HgS) and realgar (arsenic sulfide, As_4S_4) were sometimes classified as upper-grade drugs. People at the time did not regard toxic minerals only as dangerous substances that had to be avoided. They believed that careful selection, grinding, washing, and heating could reduce the risks while preserving their special effects.

Turning minerals into medicine, however, was difficult. Processing methods were often secretly passed down within groups of physicians or Daoist practitioners, and written instructions were not always detailed. Without standardized manufacturing methods, medicines with the same name could differ greatly in the quality of their ingredients and the way they were prepared. Poorly produced medicines or excessive doses could cause serious side effects. Nevertheless, people did not completely abandon toxic minerals because they believed that these substances produced powerful effects that were difficult to obtain from other ingredients.

This article examines Five-Stone Powder, one of the best-known examples of such a medicine. Five-Stone Powder combined several minerals with plant-based ingredients and ground them into a powder. At one time, it was celebrated as a medicine that could cure disease, strengthen the body, and even make the mind clearer. Later, however, it came to be remembered as a dangerous drug and a symbol of the reckless behavior of aristocrats during the Six Dynasties periods. Why did people take Five-Stone Powder even though they knew it

was dangerous? Why did it remain in use for centuries before almost disappearing after the Song dynasty? To answer these questions, we must examine not only the toxicity of the medicine but also changing ideas about medicine and the development of pharmaceutical processing techniques.

1. What Was Five-Stone Powder?

The name Five-Stone Powder refers to a medicine made by mixing several kinds of "stone drugs," or mineral ingredients. Its formula differed from one text to another, but common ingredients included stalactite(鐘乳石; primarily calcium carbonate, $CaCO_3$), purple quartz(紫石英; primarily silicon dioxide, SiO_2), white quartz(白石英; silicon dioxide, SiO_2), and red halloysite(赤石脂; $Al_2Si_2O_5(OH)_4 \cdot 2H_2O$). Plant-based ingredients such as cinnamon, dried ginger, and fangfeng(防風; the root of Saposhnikovia divaricata) were sometimes added. The Essential Prescriptions of the *Golden Coffer* preserves a formula called Purple-Quartz Cold-Food Powder(紫石寒食散), which was based mainly on purple quartz. This shows that Five-Stone Powder was not simply ground mineral dust. It was a complex prescription that combined ingredients believed to have different medicinal properties.

People during the Six Dynasties period believed that Five-Stone Powder produced intense heat inside the body. If this heat spread properly, it was thought to increase energy and make the body feel light. If the heat became blocked, however, it could cause chills, pain, and other problems. To release the medicinal heat, patients walked or exercised, wore thin clothing, and sometimes bathed in cold water. Because they also ate cold foods to reduce the heat, Five-Stone Powder was known as Cold-Food Powder (Hanshi San, 寒食散), meaning "a powdered medicine taken with cold food."

A *New Account of the Tales of the World* records that He Yan, an official of the state of Wei, claimed that Five-Stone Powder not only cured illness but also made his mind bright and clear. The famous calligrapher Wang Xizhi wrote in a letter that after taking a mineral-based powdered medicine, his body felt light and he could move as though he were flying. These accounts suggest that people did not think of Five-Stone Powder simply as a recreational drug. To them, it was a powerful medicine that could cure illness, restore physical strength, and promote long life.

Careful treatment after taking the medicine was considered essential. The *History of the Southern Dynasties* tells the story of the physician Xu Sibai, who treated Fang Baiyu after the patient had repeatedly taken Five-Stone Powder without experiencing its expected effects. Fang felt so cold that he wore several layers of clothing even during the hottest part of summer. Xu believed that heat was trapped inside the patient's body and had to be released through stimulation with cold water. When winter arrived and ice formed, he repeatedly poured cold water over Fang. The patient could not move for some time, but heat eventually began to rise from his back. According to the story, his body then became warm, and he recovered his health.

From a modern perspective, this was an extremely dangerous treatment. It is also difficult to accept every detail of the story as historical fact. Nevertheless, the account clearly demonstrates the medical reasoning of the time. The heat of Five-Stone Powder had to be properly "activated," and the physician's role was to draw out and control this heat. The

medicine was not believed to work automatically as soon as it was swallowed. Its effects depended on the patient's physical condition and on the careful management of food, clothing, bathing, and exercise.

2. How Did It Become Known as a "Drug of Decadence"?

Today, Five-Stone Powder is often described as a drug that encouraged the strange behavior and reckless lifestyles of aristocrats during the Six Dynasties periods. One of the people who helped spread this image was Lu Xun (魯迅, 1881–1936), a leading figure in modern Chinese literature. In his 1927 lecture, "The Wei-Jin Style and the Relationship between Literature, Medicine, and Wine," he argued that the unusual lifestyles of intellectuals described in *A New Account of the Tales of the World* were connected to the use of Five-Stone Powder. It later became widely regarded as a symbol of the unscientific character of premodern society and the decadence of aristocratic culture.

Lu Xun's interpretation was also connected to the conditions of modern China. Chinese intellectuals were deeply concerned about the devastating social effects of opium addiction, which they regarded as one of the main factors that had reduced China to a semi-colonial state under Western powers since the nineteenth century. When they critically examined the past, some compared Five-Stone Powder to an ancient form of opium. They connected the political disorder of the Six Dynasties period and the unconventional behavior of intellectuals with the effects of drugs.

This explanation, however, leaves an important question unanswered. If Five-Stone Powder only harmed people, why did it remain in use for centuries? Were physicians and patients of the time simply unable to recognize the danger? Historical sources show that they were well aware of its side effects. They continued to use it because they also believed it had powerful therapeutic effects. They thought that correct use and careful management could reduce its dangers.

Explaining Five-Stone Powder only as a "drug of decadence" therefore prevents us from understanding why people chose to take it and how medical knowledge changed over time. Historical research should not simply judge the past according to present-day standards. It should examine how people made decisions based on the knowledge and experience available to them and how those decisions changed over time.

3. From Blaming the Patient to Examining the Medicine

The risks of Five-Stone Powder had been recognized from an early period. In the fifth century, the physician Qin Chengzu praised it as an excellent formula that could support life. At the same time, he warned that just as water can carry a boat but also overturn it, Five-Stone Powder could either save a life or destroy it. He argued that patients could die very quickly if a physician trusted the medicine's effects without examining the patient's condition. Instead of following a single rule in every case, physicians had to adjust the prescription according to each person's body and symptoms.

During this period, side effects were mainly blamed on incorrect use and poor management after taking the medicine. Ancient medical records quoted in the Japanese medical collection, the *Prescriptions from the Heart of Medicine* (醫心方, *Ishinpō*), explain

that Five-Stone Powder was originally a good medicine and that the problem lay with the user, not the drug itself. These records provided detailed instructions, such as using different bathing methods depending on whether the patient felt hot or cold. If a particular mineral caused a headache, the patient was advised to drink warm wine. The main goal was not to reject the medicine but to observe and control how its effects moved through the body.
The Western Jin(266-317) scholar Huangfu Mi also wrote about the side effects of Five-Stone Powder. He said that he took the medicine to treat numbness but failed to follow the proper rules. As a result, he suffered so severely that he had to expose his body and eat ice in the middle of winter. Huangfu Mi likewise believed that the problem did not lie in Five-Stone Powder itself, but that its side effects resulted from improper use. However, his case cannot be accepted uncritically as evidence of the medicine's deadly dangers. Above all, the statement was a political document written to politely decline Emperor Wu's repeated orders that he enter government service. The illness and suffering described in it may therefore have reflected Huangfu Mi's actual experience, but we must also consider the possibility that his account involved some rhetorical exaggeration and political calculation intended to help him avoid official service. Physicians of the Six Dynasties periods generally understood the side effects of Five-Stone Powder as consequences of incorrect use or failure to follow the proper regimen after taking it. In other words, they attributed these side effects not to the medicine itself but to the behavior of the person taking it. Accordingly, they believed that the solution lay in taking the medicine correctly and strictly following the prescribed regimen.
By the Sui and Tang dynasties, medical explanations began to change. Earlier accounts often emphasized the patient's failure to follow the rules. Later physicians increasingly examined how the heat and toxicity of minerals affected the body itself. In the *Treatise on the Origins and Symptoms of Diseases*, the Sui physician Chao Yuanfang explained that the long-term use of Five-Stone Powder could cause a condition known as "internal wasting disease (neixiao bing, 内消病)." He believed that the heat of minerals could gather in the kidneys, leading to excessive urination and extreme weakness. He also argued that as people aged and became physically weaker, they might no longer be able to control the heat of minerals taken during their youth. This could cause new illnesses to appear many years later.

This was an important change. Side effects were no longer explained only as the result of the patient's carelessness or personal failure. Physicians attempted to connect the properties of a medicine with the body's response. They also classified people according to their pulse and physical condition and observed signs such as appetite, skin color, itching, chills, and sleepiness. Their theories and methods differed from those of modern medicine. Nevertheless, their attempts to classify symptoms and identify causes became increasingly detailed.

4. Sun Simiao's Warning: Five-Stone Powder Was Not the Same as Processed Mineral Medicine

The famous Tang(618-907) physician Sun Simiao strongly criticized Five-Stone Powder. In the *Essential Prescriptions Worth a Thousand in Gold*, he wrote that it would be better to eat a poisonous plant than to take Five-Stone Powder. He even advised readers to burn the prescription if they found it. His warning was based on his experience of seeing

several officials suffer from serious side effects. This statement may make it seem that Sun opposed every kind of mineral medicine. In the same book, however, he explained that even minerals used as ingredients in Five-Stone Powder could be made into mineral medicines that warmed the body, strengthened the bones and marrow, and prevented various diseases, provided that they were carefully selected and thoroughly processed. He instructed practitioners to examine the place of origin, color, transparency, and patterns of minerals so that they could avoid poor-quality ingredients. He also emphasized that all mineral drugs had to be thoroughly heated and refined before use. Sun therefore did not reject mineral medicine as a whole. What he rejected was the Five-Stone Powder formula, which combined several dangerous substances and strongly stimulated the body.

The *Essential Prescriptions Worth a Thousand in Gold* also records the case of Li Wenbo, an official who took white quartz(silicon dioxide, SiO_2) for a long time. After developing unusual physical symptoms, he experienced severe thirst and frequent urination. Although he received several treatments, his condition continued to worsen, and he eventually died. Sun used this case to warn that long-term use of mineral medicines could cause rare but frightening diseases. He accepted the therapeutic value of minerals but insisted that practitioners carefully control the quality of the ingredients, the method of processing, the dosage, and the length of treatment.

Sun's position was not contradictory. He distinguished dangerous compound formulas from carefully processed individual mineral ingredients. Instead of seeking the intense effects of Five-Stone Powder, he believed that it was safer to select only the necessary minerals, process them thoroughly, and use them according to the needs of each patient. This distinction became an important turning point in the decline of Five-Stone Powder and the development of new mineral-processing techniques.

5. From Enduring a Medicine after Taking It to Making It Safer before Use

Early treatments involving Five-Stone Powder focused on managing the heat and chills that appeared after taking it. Food, bathing, clothing, and exercise were used to control the patient's reaction. Over time, however, medical attention shifted toward the preparation of ingredients before they entered the body. Physicians believed that side effects could be reduced by removing dangerous materials in advance, making particles more even, and changing the properties of the ingredients. The processes of washing, cutting, grinding, boiling, roasting, and heating medicinal substances are known as paozhi(炮炙), or pharmaceutical processing.

Methods of processing mineral medicines were not systematic at first. Early medical texts often described the effects of minerals in detail but said relatively little about how to reduce their toxicity. By the Sui and Tang periods, the *Master Lei's Treatise on Pharmaceutical Processing* (雷公炮炙論, *Leigong Paozhi Lun*), an important work on pharmaceutical processing, described a complicated method for preparing cinnabar(mercury sulfide, HgS). It involved washing and crushing the mineral, grinding it for an extended period, boiling it with several plants, washing and drying it again, and finally heating it in a fire. The scientific principles behind this process were not clearly explained. Nevertheless, it shows that people increasingly believed medicinal substances should undergo several stages

of treatment rather than being used in their raw state.

Sun Simiao also instructed practitioners to crush mineral ingredients into pieces about the size of rice grains. Stones such as stalactite were to be ground with water for three days and nights until they became extremely fine. Cinnabar and realgar(arsenic sulfide, As_4S_4) were ground into powder and mixed evenly into medicinal liquids. This demonstrates an effort to control even the size of particles and the consistency of mixtures. As these methods were written down in medical books, they could be shared beyond the personal experience of individual physicians.

A further transformation took place during the Song dynasty. The government operated official pharmacies responsible for purchasing, manufacturing, and selling medicines. It also compiled standardized collections of prescriptions. The "General Guide to the Use of Medicines" by Xu Hong, included in the *Formulary of the Pharmacy Service for Benefiting the People in the Taiping Era*, emphasized the need to distinguish genuine ingredients from false ones, identify their places of origin, process them according to established methods, and measure their quantities precisely. Manufacturing practices that had depended largely on personal judgment were beginning to become shared rules managed by the state.

For example, cinnabar, realgar, and orpiment(雌黄, As_2S_3) were first crushed and finely ground. They were then processed through water levigation(shuifeifa, 水飛法), in which only the finest particles suspended in water were collected. Red halloysite($Al_2Si_2O_5(OH)_4 \cdot 2H_2O$) and white halloysite(白石脂) were heated until they glowed, cooled, and ground into fine powder again. Plant and animal ingredients were also carefully processed. The central core of Polygala root(yuanzhi, 遠志), a medicinal herb used to treat insomnia, anxiety, forgetfulness, palpitations, and coughs accompanied by phlegm, was removed because it was believed to cause discomfort. Pinellia tuber(banxia, 半夏), a medicinal herb used to reduce phlegm, relieve coughing and vomiting, and treat sensations of fullness or discomfort in the stomach, was washed seven times in hot water to reduce its irritating effect on the throat. The poisonous blister beetle(banmao, 斑蝥), a highly toxic medicinal substance traditionally used to remove warts, boils, and malignant masses, was processed by removing its wings and legs and roasting it with glutinous rice. In other words, different preparation methods were developed to reduce the risks of each ingredient while preserving its useful effects.

The central change was a shift in emphasis from "managing the drug's effects after the patient has taken it" to "controlling the properties of ingredients before the patient takes them." The secret knowledge of individual physicians was recorded in texts, while different manufacturing methods were compared and standardized. As a result, there was less need for a medicine like Five-Stone Powder, whose users had to endure extreme heat and cold to obtain its supposed benefits.

6. Why Did Five-Stone Powder Disappear?

From the late Tang period onward, historical accounts of the misuse and side effects of Five-Stone Powder became less frequent. After the Song dynasty, they almost disappeared. This does not mean that people suddenly decided that all mineral medicines were dangerous

and prohibited them. Minerals continued to be used as medicine, but prescriptions that combined several powerful ingredients in the manner of Five-Stone Powder became less common.

The Northern Song scholar Shen Kuo examined the principles of Five-Stone Powder in the *Brush Talks from Dream Brook*. He observed that descriptions in older medical texts appeared to contradict one another. Even when he asked experienced physicians of his own time, no one could properly explain the reason for these contradictions. This suggests that by the eleventh century, Five-Stone Powder had already moved away from actual medical practice, and knowledge about the principles behind its formula had not been fully transmitted. Shen emphasized that ancient books should not be accepted without question. Instead, physicians had to examine the logic of a prescription and its actual effects.
Several factors contributed to the decline of Five-Stone Powder. First, it was highly toxic and caused serious side effects. Second, because it required several expensive minerals, it was mainly available to members of the upper class. Third, patients had to carefully regulate their diet, bathing, clothing, and physical activity after taking it, since even a small mistake could prove fatal. Although these three factors played important roles in its decline, the fourth brought about a more fundamental change: during the Tang and Song dynasties, increasingly advanced techniques for selecting, refining, and combining individual mineral medicines made it possible to preserve their therapeutic benefits while reducing their toxicity. This development gradually made complex and dangerous formulas such as Five-Stone Powder less necessary. If physicians could achieve similar therapeutic effects in a safer way, there was little reason to continue using a dangerous compound formula.

The disappearance of Five-Stone Powder should not be understood simply as the result of medical failure or a sudden recognition of its dangers. Mineral medicine itself did not disappear. Instead, some of the therapeutic benefits associated with Five-Stone Powder were preserved through improved techniques for refining and combining individual minerals. What gradually vanished was the older medical practice of relying on the patient's lifestyle and physical endurance to control toxicity after taking the medicine. As safer medicines offering similar therapeutic effects became available, Five-Stone Powder gave way to more advanced pharmaceutical-processing techniques and a more systematic approach to medical treatment.

Conclusion: What the History of One Medicine Tells Us about Scientific Change

Five-Stone Powder was a highly toxic medicine that could cause deadly side effects. Yet people in medieval China also regarded it as a powerful treatment and health-enhancing drug whose effects were difficult to replace. For this reason, the first question asked by medieval Chinese physicians was not "Should we eliminate this medicine?" but "How can we preserve its benefits while reducing its side effects?"

The answer changed over time. During the Six Dynasties period, physicians emphasized correct use and careful management after taking the medicine. During the Sui and Tang dynasties, they developed more detailed explanations of diseases caused by the heat and toxicity of mineral drugs. They also began to distinguish Five-Stone Powder from carefully processed mineral medicines. During the Song dynasty, pharmaceutical processing

methods were increasingly standardized through medical texts and government institutions. The responsibility for controlling danger gradually shifted away from the patient's endurance and the individual physician's judgment. Instead, risks were reduced in advance through the careful selection of ingredients and control of the manufacturing process. This shift also transformed the meaning of a "good medicine." A good medicine was no longer understood simply as a drug that produced a powerful effect. It increasingly came to mean one whose quality and dosage could be predicted and whose side effects could be controlled.

If we view Five-Stone Powder merely as a symbol of superstition or decadence, as many modernizing intellectuals did in the early twentieth century, we overlook its broader historical significance. Its rise and decline were part of a long-term effort to transform dangerous natural substances into effective and manageable medicines. This process also illustrates how medical knowledge, once held and transmitted privately, was gradually recorded, shared, and standardized. The disappearance of Five-Stone Powder, therefore, should not be understood simply as evidence of the failure of premodern medicine. Rather, it reflected advances in pharmaceutical technology and the continuing search for safer, more reliable ways to use medicinal substances.

Works Cited

Primary Sources

Treatise on the Origins and Symptoms of Diseases (Zhubing Yuanhou Lun, 諸病源候論).

History of the Southern Dynasties (Nan Shi, 南史).

A New Account of the Tales of the World (Shishuo Xinyu, 世說新語).

Master Lei's Treatise on Pharmaceutical Processing (Leigong Paozhi Lun, 雷公炮炙論).

Brush Talks from Dream Brook (Mengxi Bitan, 夢溪筆談).

Essential Prescriptions Worth a Thousand in Gold (Beiji Qianjin Yaofang, 備急千金要方).

Prescriptions from the Heart of Medicine (Ishinpō, 醫心方).

Essential Prescriptions of the Golden Coffer (Jingui Yaolüe, 金匱要略).

Formulary of the Pharmacy Service for Benefiting the People in the Taiping Era (Taiping Huimin Hejiju Fang, 太平惠民和劑局方).

Secondary Sources

Liu, Yan. Healing with Poisons: Potent Medicines in Medieval China. Seattle: University of Washington Press, 2021.

Richter, Antje, and Charles Chace. "The Trouble with Wang Xizhi: Illness and Healing in a Fourth-Century Chinese Correspondence." T'oung Pao 103 (2017).

Satō, Toshiyuki. "Wang Xizhi and Five-Stone Powder" [王羲之と五石散]. The Hiroshima University Studies, Graduate School of Letters 65 (2005).

Spiro, Audrey. Contemplating the Ancients: Aesthetic and Social Issues in Early Chinese Portraiture. Berkeley: University of California Press, 1990.