Understanding How Personal Health Characteristics Influence Healthcare Spending

Zaid Ahmed
School of Data Science
UNC CHARLOTTE
Charlotte, United States
zahmed1@charlotte.edu

Utkarsh Jain
School of Data Science
UNC CHARLOTTE
Charlotte, United States
ujain@charlotte.edu

Kyle Jeffrey
School of Data Science
UNC CHARLOTTE
Charlotte, United States
kjeffre2@charlotte.edu

Alexi Grafakos
School of Data Science
UNC CHARLOTTE
Charlotte, United States
agrafako@charlotte.edu

Baran Narravula
School of Data Science
UNC CHARLOTTE
Charlotte, United States
bnarravu@charlotte.edu

***Abstract*—As healthcare costs continue rising, many U.S. residents struggle to afford care without compromising quality. While it's understood that personal health characteristics influence health outcomes, their specific financial impact remains underexplored. This study quantifies how physical and mental health metrics, behavioral habits, and chronic conditions affect healthcare expenditures. Using a data-driven approach, we classified individuals into distinct health clusters based on comprehensive metrics including BMI, depression scores, physical activity levels, and chronic disease status. Our analysis revealed striking financial disparities—individuals in the Low to Mid-Tier Health cluster averaged $28,707 in annual healthcare costs compared to just $8,509 for those in the Mid to High-Tier Health cluster. Through multiple regression modeling, we identified specific factors most strongly associated with increased spending: elevated depression scores, higher BMI, reduced physical activity, poor dietary habits, inadequate sleep, and the presence of chronic conditions such as heart disease and cancer. Our neural network model achieved 98.85% accuracy in validating our health cluster classifications in Phase One using NHANES data, confirming that these clusters effectively captured health status based on objective metrics rather than relying solely on self-reported assessments. In Phase Two, we identified the specific health characteristics that most significantly influence healthcare expenditures, going beyond the simple conclusion that poorer health correlates with higher spending. These findings emphasize the significant financial impact of modifiable health factors and highlight the need for targeted policies/and or recommendations that address these specific cost drivers to reduce the healthcare financial burden on individuals and systems.**

***Index Terms*—healthcare costs, preventive health, regression analysis, lifestyle impact**

I. Introduction

Annually, the U.S. spends over $4.3 trillion on healthcare, with 90% of these costs resulting from chronic and preventable diseases. As healthcare expenses rise, many individuals turn to clinical interventions such as medication and surgery, believing these treatments will improve health. However, while clinical approaches can improve health, they often come with a high price tag. This drives our research question: How do specific personal health characteristics and lifestyle factors influence healthcare spending? Our study investigates the relationship between various health attributes—including physical metrics, mental health indicators, behavioral habits, and chronic conditions—and their association with healthcare expenditures. By identifying which personal health characteristics most significantly impact healthcare costs, we aim to uncover the key drivers behind the financial burden of medical care in the U.S. This approach will provide data-driven insights that can help inform individuals on specific factors that directly lead to an increase in healthcare spending.

II. Background

Physical health serves as the primary indicator for evaluating overall health status and healthcare expenditures. Specifically, obesity, cardiovascular health, and lifestyle habits are strong predictors of cumulative healthcare expenditure over time. Body Mass Index (BMI), a widely accepted measure for obesity, strongly correlates with increased risk of diabetes, heart disease, and other chronic conditions. According to the CDC (2022), obesity affects approximately 40.3% of U.S. adults and 19.7% of children aged 2-19. These high obesity rates result in substantial healthcare costs through both direct expenses, such as medications and surgical interventions, and indirect costs, including loss of productivity and disability claims [1].

Similarly, poor cardiovascular health significantly increases medical expenditures. The Chicago Heart Association Detection Project emphasizes that early preventive measures for cardiovascular diseases substantially reduce healthcare costs over time [2]. However, adherence to essential preventive recommendations such as regular exercise and balanced nutrition remains low. The CDC (2022) reported that only 28% of men and 20% of women meet recommended physical activity guidelines, a deficit which subsequently elevates cardiovascular disease risk and consequently increases associated healthcare costs [2].

Mental health status represents another critical yet frequently overlooked characteristic impacting healthcare expenditures. According to the West Health/Gallup Survey (2024), 75% of U.S. adults deemed mental health treatment to be of less importance compared to physical healthcare, despite mental health's significant role in chronic disease development and associated medical costs [3]. The CDC reported 57.2 million physician visits in 2019 that were attributed primarily to mental health disorders [4]. Depression, anxiety, and other mental health disorders raise overall healthcare costs in several ways: they generate expenses for treatment, often aggravate physical illnesses, and drive up the cost of treating those worsened physical conditions. Approximately 28% of Americans forgo or delay addressing mental health issues, which subsequently exacerbates existing health conditions, leading to higher healthcare costs as individuals ultimately require more intensive interventions [5]. Mental health metrics such as depression scores and psychological well-being assessments represent critical variables when examining the relationship between personal health characteristics and healthcare spending patterns. The quantifiable impact of mental health status on healthcare expenditures highlights the importance of including these variables in comprehensive analyses of healthcare cost drivers.

Despite available preventive and cost-effective interventions, broader systemic issues exacerbate healthcare expenses in the U.S. According to the American Hospital Association (2024), Americans currently hold at least $220 billion in medical debt, primarily due to inadequate health insurance coverage, high deductibles, and complex policies [6]. Systemic inefficiencies, including a lack of universal healthcare, ineffective payment structures, and inadequate preventive care, further inflate these costs. The U.S. allocates approximately 17.6% of its GDP to healthcare, considerably higher than the EU's 10.4%, without any better health outcomes, highlighting profound inefficiencies in healthcare delivery[7]. Insurance gaps compound these issues; around 27 million Americans are uninsured, significantly limiting access to necessary medical care and leaving individuals at risk of financial ruin if faced with a medical emergency. Furthermore, under insured individuals face substantial out-of-pocket expenses that often deter preventive care and medical recommendations [8].

A study published in *Health Affairs* found that healthcare spending has risen from 1987 - 2011, with 77.6 % of the increase attributed to individuals managing four or more chronic conditions [9]. Additionally, the study quantifies how 11.4% - 23.5% of the growth in the spending for major chronic conditions such as diabetes, hypertension, and heart disease. These findings highlight the growing economic strain that chronic illnesses place on the healthcare system, setting the stage for our analysis.

Healthcare organizations often prioritize reactive treatment over preventive strategies due to misaligned financial incentives, specifically by treating symptoms rather than providing an answer to the underlying cause. Health institutions, such as pharmacies and hospitals, alongside Health Insurance companies, benefit financially by continuously prescribing symptom treatment rather than selling a complete solution. Research suggests that emphasizing preventive measures could substantially decrease hospitalizations and emergency care, alleviating financial strain on individuals and the healthcare system [7].

III. Methodology

Our research aims to provide real-world implications and comprehensively analyze healthcare expenditures. All data was de-identified and sources were taken from public repositories, eliminating risks to participant confidentiality. We employ a comprehensive, quantitative, data-driven approach utilizing the National Health and Nutrition Examination Survey (NHANES) with 16,957 samples and the Medical Expenditure Panel Survey (MEPS) with 5,218 samples. NHANES provides individual-level health and behavioral metrics such as Body Mass Index (BMI), physical activity, coronary heart disease, diet health, and many others. MEPS contributes detailed data on mental health indicators (e.g., feelings of hopelessness or depression), as well as chronic illnesses such as asthma, high blood pressure, cancer, and diabetes. It also provides insight into healthcare expenditures, including out-of-pocket expenses, total charges, and insurance coverage from both private and public sources.

Our analysis consists of two phases: Phase One focuses on identifying the health status of individuals based on physical, mental, and behavioral health metrics alongside various stressors and lifestyle habits, while Phase Two focuses on how those factors influence healthcare spending. Phase One utilizes our NHANES data while Phase Two utilizes our MEPS data. The purpose of Phase One is to establish objective health classifications based on measurable health metrics rather than relying on potentially biased self-reported health assessments. Understanding individuals' health levels is crucial for us as it establishes a methodological foundation for our subsequent analysis of healthcare spending in Phase Two.

We clean both datasets extensively to ensure rigorous data interpretability before modeling. Due to most of our data falling within the ordinal and categorical nature, we standardize and one-hot encode many variables to ensure robust calculations throughout both modeling phases. We conduct hierarchical clustering for both phases to group individuals into distinct health-based clusters. By integrating dendrogram plots into our hierarchical clustering methodologies, we can statistically assess the optimal number of clusters properly suited for our data. We then validate our clusters using ANOVA and Chi-Square tests to confirm statistical variation between clusters and confirm cluster label associations with self-reported health status. For additional validation in Phase One only, we employ a feed-forward neural network using lifestyle habits and stressors as inputs to predict NHANES cluster labels, ensuring our assigned health classifications to the clusters represent objective health status rather than merely reflecting our self-assigned bias.

It's important to note that our research methodology was specifically designed to identify the individual health characteristics that drive healthcare spending, rather than simply demonstrating that overall poor health leads to higher costs. The neural network validation was conducted only in Phase One with NHANES data to establish reliable health cluster classifications. In Phase Two, we applied our clustering approach to MEPS data and then used regression models and gradient boosting trees to isolate specific health factors that influence spending patterns, regardless of overall health status. The regression models consist of two OLS regression models and an XGBoost tree to assess general patterns, relationships, and associations between what personal health characteristics lead to certain health levels and how they influence healthcare spending. They will be specifically designed to isolate individual health characteristics that influence healthcare expenditures, rather than simply confirming that poorer overall health leads to higher costs. By including interaction terms and controlling for multiple variables simultaneously, we aim to identify which specific health factors, such as depression in younger adults or comorbid conditions, drive spending patterns independently of general health status. This approach provides more nuanced and actionable insights than broad health categorizations alone. Overall, our methodology focuses on identifying attributes that influence health and examining how these attributes affect healthcare spending, regardless of one's health status.

IV. Results

The first phase of our analysis involved clustering individuals based on a comprehensive set of health indicators from the NHANES dataset, including diagnosed diseases, mental health metrics, and behavioral practices. Our goal was to establish health classifications based on objective metrics rather than relying solely on subjective self-assessments. Our hierarchical clustering dendrogram revealed three distinct clusters that adequately captured the variance in our data. We validated these clusters through multiple statistical approaches: ANOVA tests confirmed significant differences between clusters across all variables (p values less than 0.05), while chi-square analysis verified the association between cluster membership and self-reported general health ratings (p value less than 0.05). These complementary statistical validations supported our clustering approach and informed our classification of individuals into Poor Health for Cluster 1, Good Health for Cluster 2, and Average Health for Cluster 3.

Cluster analysis identified three distinct phenotypes with markedly different cardio-metabolic and behavioral profiles. Cluster 1 (Poor Health) shows the highest cardio metabolic risk: a mean BMI of 32.09 kg/m² (obese range) and the second-highest weekly sedentary time (376.26 min). It also has the worst prevalence of chronic disease, with the lowest CHD index (0.89) and HBP index (0.30), together with the highest diabetes prevalence (mean = 1.00 on a 1 = Yes, 2 = No, 3 = Maybe scale). On lifestyle measures, this group reports only fair diet quality (mean = 3.08 on a 1 = Excellent to 4 = Fair scale) and the second lowest depressive symptoms (0.23 on a 0 = None to 3 = Nearly Every Day scale). Cluster 2 (Good Health) consistently fares best: lowest BMI (29.08 kg/m²), least sedentary time (355.83 min/week), highest CHD (0.97) and HBP (0.68) indices, lowest diabetes prevalence (2.03), best diet quality (2.98), and minimal depression (0.21). Cluster 3 (Average Health) sits between the other two on most metabolic measures (BMI = 30.44 kg/m²; sedentary time = 383.13 min; CHD index = 0.94; HTN index = 0.52; diabetes = 1.80) but stands out for its psychosocial burden, exhibiting the highest depression score (2.52) and the poorest diet quality (3.38). These results demonstrate that metabolic, behavioral, and mental health indicators do not distribute randomly across individuals but instead aggregate into distinct, lifestyle-driven health phenotypes.

TABLE I: **Figure 1.** Mean health indicator values by cluster (NHANES data).

Variable	**Cluster 1** (Poor Health)	**Cluster 2** (Good Health)	**Cluster 3** (Average Health)
BMI (kg/m^2)	32.09	29.08	30.44
Inactive Time (min/week)	376.26	355.83	383.13
Coronary Heart Disease (0–1)	0.89	0.97	0.94
High Blood Pressure (0–1)	0.30	0.68	0.52
Depression (0–3)	0.23	0.21	2.52
Healthy Diet (1–4)	3.08	2.98	3.38
Diabetes (1–3)	1.00	2.03	1.80

Labels and Scales:

- **Coronary Heart Disease (0–1):** 0 = Yes, 1 = No.
- **High Blood Pressure (0–1):** 0 = Yes, 1 = No.
- **Depression (0–3):** 0 = Not at All, 1 = Sometimes, 2 = Several Days, 3 = Nearly Every Day.
- **Healthy Diet (1–4):** 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair.
- **Diabetes (1–3):** 1 = Yes, 2 = No, 3 = Somewhat.

To confirm that our health clusters represented objectively meaningful health status categories rather than merely reflecting self-reported health assessments, we developed a neural network model specifically for the NHANES dataset in Phase One. This model was designed to predict cluster membership based on all health-related variables except the self-reported general health variable. The architecture consisted of two hidden layers with 64 and 32 nodes, and dropout regularization was applied to prevent overfitting across 100 training epochs. The model achieved a remarkable 98.85% accuracy on the validation dataset, with a very low loss of 0.0277, demonstrating that our clusters captured coherent, predictable patterns of health characteristics. This validation confirmed that our ”Poor”, ”Good”, and ”Average” cluster classifications represented meaningful distinctions based on comprehensive health metrics rather than subjective self-assessments alone. Having validated this clustering approach in Phase One, we proceeded to Phase Two without repeating the neural network validation, applying our proven clustering methodology to the MEPS dataset.This helped in examining the relationship between specific health characteristics and healthcare expenditures.

In phase two, we performed an additional hierarchical clustering analysis to examine differences in healthcare spending, biological factors, personal habits, and disease stressors, expanding upon the initial attributes identified in phase one. The result from our dendrogram revealed two adequate clusters to capture the total variance within the data. Our ANOVA tests yielded statistically significant results across all variables with p-values well below the 0.05 significance level, yet again indicating our clusters are statistically differentiated. To assign health labels to the two clusters identified in phase two, we examined key demographic and health-related variables. Clear patterns emerged that differentiated the clusters. Cluster 1

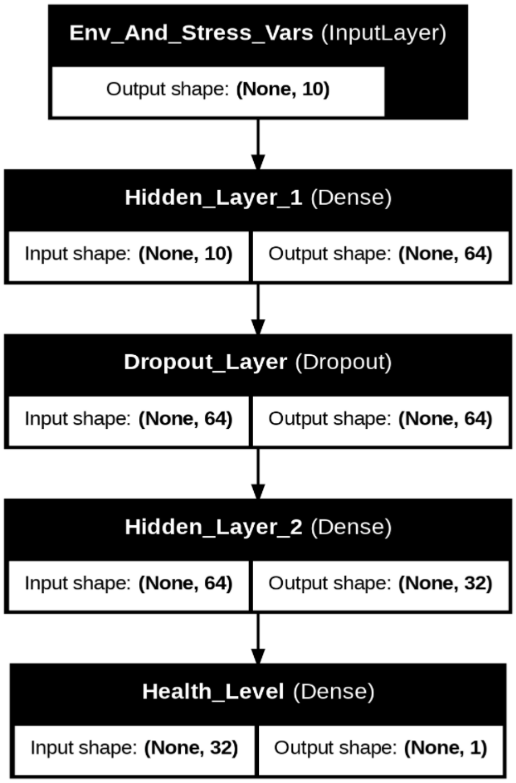

Figure 2. Neural network architecture for health classification

(later labeled as Low to Mid-Tier Health) showed significantly worse outcomes across multiple health dimensions. On average, individuals in this cluster were slightly older (63.9 vs. 61.2 years) and reported considerably poorer psychological well-being, with an average Psychological Distress Score of 9.4 compared to 1.9 in Cluster 2 (on a scale where 11 indicates poor mental health).

Cluster 1 also exhibited more severe physical and social limitations. Their average score for Health Limits Work was 3.3 versus 1.7 in Cluster 2 (1 = no limitation, 5 = severe limitation). Similarly, Little Interest in Activities averaged 1.5 in Cluster 1 compared to just 0.16 in Cluster 2 (on a 0–3 scale where 3 indicates no interest at all), and Health Stopped Social Activity was 2.9 versus 4.5 (on a 1–5 scale, where higher values reflect better social functioning). Depression levels were notably higher in Cluster 1, with a Felt Depressed score of 1.4 compared to 0.1 in Cluster 2 (on a scale of 0-3, where 3 indicates severe depression). In addition to these psychosocial differences, Cluster 1 showed a consistently higher prevalence of chronic diseases, including arthritis, asthma, hypertension, cancer, coronary heart disease, diabetes, and high cholesterol. These conditions were coded as 1 = yes (diagnosed) and 2 = no (not diagnosed), and Cluster 1's average values were 0.2 lower, indicating a higher diagnosis rate across the board.

Given the consistent disadvantages across mental, physical, and chronic disease indicators, we labeled Cluster 1 as representing Low to Mid-Tier Health and Cluster 2 as Mid to High-Tier Health. This categorization provides a solid foundation for analyzing variations in healthcare spending across health status groups.

TABLE II: **Figure 3.** Mean variable values by cluster — MEPS data.

Variable	**Cluster 1**	**Cluster 2**
	Low to Mid-Tier Health	*Mid to High-Tier Health*
Age (in years)	63.85	61.98
Psychological distress score	9.40	1.91
Health limits work (1–5)	3.33	1.74
Little interest in activities (0–3)	1.53	0.17
Health stopped social activity (1–5)	2.87	4.52
Felt depressed (0–3)	1.41	0.11
Arthritis diagnosis (0–1)	0.32	0.56
Asthma diagnosis (0–1)	0.73	0.86
Hypertension diagnosis (0–1)	0.11	0.17
Cancer diagnosis (0–1)	0.75	0.80
Coronary heart disease (0–1)	0.79	0.89
Diabetes diagnosis (0–1)	0.61	0.75
High cholesterol diagnosis (0–1)	0.32	0.41
Total health care expenditure ($)	**28 706.98**	**8 509.34**

Notes:

- *Psychological distress score (0–24):* 0 = no limitation, 24 = severe limitation.
- *Health Limits Work (1–5):* 1 = no limitation, 5 = severe limitation.
- *Little Interest in Activities (0–3):* 0 = always interested, 3 = no interest at all.
- *Health Stopped Social Activity (1–5):* 1 = health always prevents socializing,
  5 = health never prevents socializing.
- *Felt Depressed (0–3):* 0 = not at all, 3 = nearly every day.
- *Chronic Disease Indicators (0–1):* 0 = diagnosed (yes), 1 = not diagnosed (no).
- *Total Health Care Expenditure ($):* Annual per-person spending.

Notably, some attributes used in Phase One were also carried forward into Phase Two. Our analysis revealed that individuals in the Low to Mid-Tier Health cluster incurred significantly higher annual healthcare expenditures, averaging $28,707, compared to $8,509 for those in the Mid to High-Tier Health cluster, as illustrated above in Figure 3a.

To explore the causal relationship between health factors and healthcare spending, we used two multiple OLS regression models in our main analysis. As such, our base OLS model, which does not include interaction, quadratic, or log-transformed terms, performed well. The insights gleaned highlighted several factors affecting total healthcare spending, which closely align with individuals' health status. Our analysis revealed that negative lifestyle factors—such as increased depression, poorer psychological health, and reduced interest in physical activity—were significantly associated with higher healthcare spending across various forms of insurance coverage and out-of-pocket expenses, as shown in Figure 5. These associations demonstrated strong statistical significance, with p-values for variable coefficients well below 0.05. Additionally, stress-related health conditions, including heart disease, cancer, diabetes, arthritis, asthma, high blood pressure, and high cholesterol, also showed strong associations with increased healthcare spending, again with robust statistical significance (p-values well below 0.05). However, the model posed an R-squared value of around 50%, which signifies that while the model can be relied on to derive suitable patterns and insights, such a model cannot adhere to precise inference. Nonetheless, drawing statistical insights from phase 1, we hypothesize that individuals exhibiting behaviors such as higher smoking rates, elevated BMI, poor dietary habits, lower sleep quality, reduced physical activity, or conditions like diabetes, depression, and heart disease, tend to allocate substantially more of their income toward healthcare expenditures.

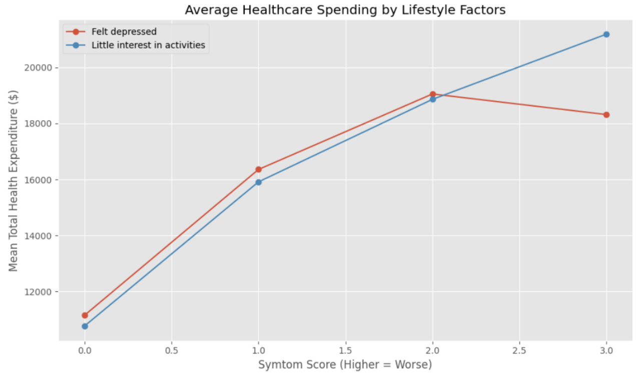

Figure 4. Impact of lifestyle factors on health care expenditure

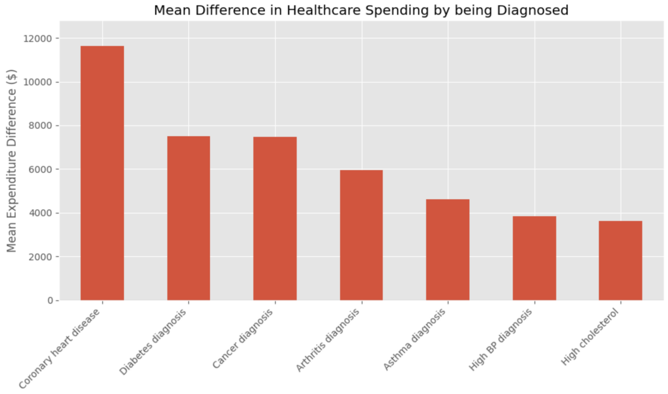

Figure 5. Mean Difference in Healthcare Spending by Diagnosis

To identify more nuanced spending patterns, we developed a specialized regression model using log-transformed healthcare expenditure as the dependent variable. We implemented various interaction terms to decipher the joint impact of multiple variables simultaneously while holding all else constant. This approach improves interpretability by expressing relationships in percentage terms and revealing how combinations of health factors collectively influence spending beyond their individual effects. This enhanced model maintained strong statistical validity, with a high F-statistic and statistically significant p-values for key health characteristics including depression scores, psychological distress metrics, chronic disease diagnoses (particularly coronary heart disease and asthma), and physical activity limitations.

The impacts on healthcare spending were substantial: individuals reporting work limitations due to physical conditions spent approximately 17% more on healthcare than those without such limitations. Similarly, those reporting general activity restrictions due to physical conditions spent approximately 17% more. Age combined with mental health status revealed particularly striking patterns, with individuals aged 40 and younger who reported depression spending approximately 33% more on healthcare than their non-depressed counterparts. The model revealed additional interaction effects, with individuals diagnosed with both diabetes and depression spending approximately 5% more on healthcare, highlighting how the combination of physical and mental health conditions can compound spending impacts.

Most notably, the absence of chronic conditions was associated with significant cost reductions: individuals without diagnoses of cancer, hypertension, or heart disease spent approximately 10-30% less on healthcare than those with these conditions. For example, individuals without both hypertension and high cholesterol spent 12.5% less on healthcare than those with these conditions. While this enhanced model's explanatory power (R-squared of around 40%) indicates it remains more suitable for identifying patterns than for precise individual predictions, we implemented robust standard errors to address potential heteroskedasticity and ensure the statistical reliability of our coefficient estimates.

Building on our Phase Two clustering and regression analyses, we next trained an XGBoost model to predict annual healthcare spending from the full suite of health indicators. The feature-importance plot (Figure 6) reveals that functional limitations overwhelmingly drive spending: "Health limits activities" alone accounts for roughly 30% of the model's explanatory power, with "Health limits work" contributing another 18%. Classic chronic diseases—diabetes (around 11%), arthritis (around 8%), and coronary heart disease (around 7%)—follow in importance, while other conditions (asthma, cancer) and psychosocial measures (social-activity limitations, age, high cholesterol, hypertension, psychological distress, depression, and lack of interest in activities) each account for only 1–5% of the total. Although these predictors capture coherent patterns, the model's RMSE ($8 000) remains large compared to the mean spending ($9 000), indicating that unmeasured clinical events, socioeconomic factors, or random outliers still drive much of the residual variation. Practically, these findings suggest that interventions aimed at preserving or restoring physical function could yield the greatest reductions in healthcare costs, while chronic-disease management and psychosocial support play important but secondary roles.

V. Discussion

Our analysis provides compelling evidence that specific personal health characteristics significantly influence healthcare expenditures by directly impacting utilization patterns

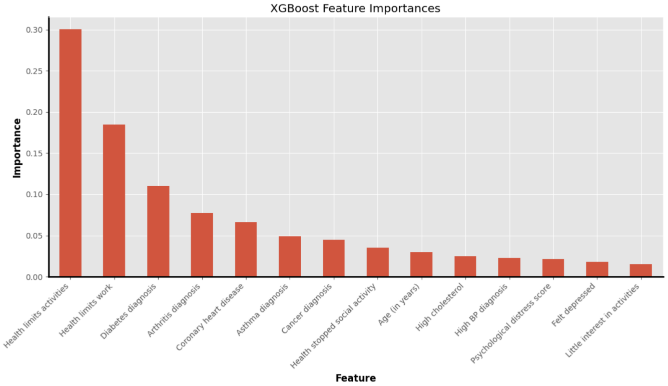

Figure 6. Feature Importance

and treatment needs. By integrating hierarchical clustering, regression modeling, and predictive analysis, we successfully delineated distinct health-based clusters and uncovered meaningful associations between individual health attributes and healthcare spending. The core aim of our research was to identify and quantify how physical metrics, mental health indicators, behavioral habits, and chronic conditions affect healthcare costs, and our results demonstrate clear and substantial relationships.

It is important to emphasize that our research objective extended beyond simply demonstrating that poor overall health leads to higher healthcare costs—a relatively intuitive conclusion. Instead, we focused on identifying and quantifying the specific health characteristics and their interactions that most significantly drive healthcare expenditures. This nuanced approach allowed us to isolate individual factors—such as depression in younger adults or the combination of multiple chronic conditions—that substantially impact healthcare spending, providing more actionable insights than broad health categorizations alone. The regression analyses, particularly our enhanced model with interaction terms, were critical in establishing these specific relationships between health characteristics and spending patterns.

The regression analyses provided deeper insights into specific factors significantly driving healthcare costs. Notably, psychological health indicators emerged as powerful predictors, with individuals aged 40 and younger who reported depression spending approximately 33% more annually on healthcare than their non-depressed counterparts. Physical limitations showed similar financial impacts, with those reporting work or activity restrictions spending approximately 17% more annually on healthcare. The interaction of multiple health conditions created compounding effects, as evidenced by individuals diagnosed with both diabetes and depression spending 5% more annually than those with single conditions. These findings demonstrate how personal health characteristics interact in complex ways to influence healthcare expenditure patterns.

Our analysis of chronic condition impacts revealed that the absence of chronic diseases was associated with substantial cost reductions, with individuals without diagnoses of cancer, hypertension, or heart disease spending approximately 10-30% less annually on healthcare than those with these conditions. This quantifiable relationship between disease status and healthcare spending reinforces the financial significance of preventing and effectively managing chronic conditions. The neural network model's remarkable 98.85% accuracy in classifying health status based on common metrics further validates the robust connection between measurable health characteristics and overall health profiles.

Our XGBoost prediction model revealed critical insights about the relative importance of various health factors in driving healthcare spending. Most notably, functional limitations emerged as the dominant predictors, with ”Health limits activities” accounting for approximately 30% of the model's explanatory power and ”Health limits work” contributing another 18%. This finding suggests that interventions aimed at preserving or restoring physical function could yield the greatest reductions in healthcare costs. Traditional chronic diseases—diabetes (11%), arthritis (8%), and coronary heart disease (7%)—followed in importance, while psychosocial measures each accounted for only 1-5% of the total. Despite capturing these coherent patterns, the model's RMSE of $8,000 relative to mean spending of $9,000 indicates substantial unexplained variation, likely attributable to unmeasured clinical events, socioeconomic factors, or random outliers.

Interpreted within the broader context presented in our background review, these findings align closely with established research highlighting obesity, cardiovascular conditions, and mental health disorders as major drivers of healthcare costs in the U.S. Indeed, our results confirm that BMI, physical inactivity, psychological distress, and chronic disease diagnoses are all significantly associated with healthcare spending. The identification of specific percentage impacts for various health characteristics provides quantifiable evidence of their economic significance, addressing a key gap in previous research that had established health-cost relationships without precise financial quantification.

The comprehensive health profiles identified in our initial cluster analysis demonstrate how multiple health characteristics tend to accumulate within individuals, creating distinct patterns with significant financial implications. While our initial NHANES analysis revealed three distinct health clusters with varying cardio-metabolic, behavioral, and mental health profiles, our subsequent MEPS analysis consolidated to two clusters with clear financial implications. The Low to Mid-Tier Health cluster exhibited interconnected challenges—obesity, physical inactivity, poor mental health, and higher prevalence of chronic conditions—that collectively contributed to their substantially higher healthcare costs, averaging $28,707 compared to just $8,509 for those in the Mid to High-Tier Health cluster. This striking financial disparity supports our thesis that personal health characteristics form profiles that significantly impact healthcare expenditures.

However, it is important to acknowledge several limitations and potential biases inherent in our study. Firstly, the reliance on self-reported data from NHANES and MEPS may introduce reporting bias, particularly regarding sensitive information such as mental health status, dietary habits, and lifestyle behaviors. Respondents may under report or inaccurately recall information, affecting data reliability. Additionally, while our regression models established strong correlations between health characteristics and spending, the cross-sectional nature of our data limits our ability to establish definitive causal relationships. The improved R-squared values of our regression models (50% for the base model and 40% for the enhanced model) indicate that while these models effectively identify patterns and relationships, they still explain only a portion of the total variance in healthcare spending, consistent with the RMSE observations from our XGBoost model.

Furthermore, while hierarchical clustering effectively distinguishes discrete health groups, there remains the possibility of unobserved confounding factors influencing both cluster distinctions and spending patterns. Although we rigorously validated cluster differences through statistical methods, our analysis did not fully control real-world complexities, including socioeconomic status, geographical variations in healthcare access, and systemic healthcare disparities. These factors may independently influence both health characteristics and healthcare expenditures, potentially affecting the interpretation of our results.

Despite these limitations, our findings present a compelling case for the substantial financial impact of personal health characteristics on healthcare spending. Clarifying how specific health attributes and their interactions affect healthcare costs provides valuable insights for healthcare policy, insurance design, and individual health management. Future research should focus on longitudinal studies to establish causality more definitively, incorporate additional socioeconomic variables to control for potential confounding factors, and explore how targeted interventions addressing specific health characteristics might reduce healthcare expenditures.

Unlike studies that merely confirm the intuitive relationship between overall health status and healthcare costs, our research has identified specific, quantifiable drivers of healthcare spending. By isolating factors such as depression in younger adults (33% higher spending), physical limitations (17% higher spending), and comorbid conditions, we provide targeted insights that can inform more precise interventions and policies. Most importantly, our feature importance analysis through XGBoost modeling provides clear guidance on intervention priorities, emphasizing the fundamental role of maintaining physical function over disease-specific management strategies. These findings demonstrate that healthcare spending is influenced by complex interactions between specific health characteristics rather than just overall health status, offering a foundation for more nuanced approaches to healthcare cost management.

References

[1] N. Elgaddal, E. A. Kramarow, and C. Reuben, “Products - data briefs - number 443 - August 2022,” *Centers for Disease Control and Prevention*, 2022. [Online]. Available: https://www.cdc.gov/nchs/products/databriefs/db443.htm. Accessed: Feb. 27, 2025.

[2] W. Koyama, “Lifestyle change improves individual health and lowers healthcare costs,” *Methods of Information in Medicine*, 2018. [Online]. Available: https://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0038-1634332. Accessed: Feb. 27, 2025.

[3] D. Robeznieks, “Mental health is part of physical health. Why isn’t it treated as such?,” *AAMC*, 2024. [Online]. Available: https://www.aamc.org/news/mental-health-part-physical-health-why-isn-t-it-treated-such. Accessed: Apr. 10, 2025.

[4] National Center for Health Statistics, “National ambulatory medical care survey—2019,” *Centers for Disease Control and Prevention*, 2019. [Online]. Available: https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2019-namcs-web-tables-508.pdf. Accessed: Feb. 27, 2025.

[5] S. Rakshit *et al.*, “How does cost affect access to healthcare?,” *Peterson-KFF Health System Tracker*, 2024. [Online]. Available: https://www.healthsystemtracker.org/chart-collection/cost-affect-access-care/#Percent%20of%20adults%20who%20reported%20barriers%20to%20accessing%20medical%20care,%202022. Accessed: Feb. 27, 2025.

[6] American Hospital Association, “AHA Senate statement on what can Congress do to end the medical debt crisis in America,” *American Hospital Association*, 2024. [Online]. Available: https://www.aha.org/testimony/2024-07-10-aha-senate-statement-what-can-congress-do-end-medical-debt-crisis-america. Accessed: Feb. 27, 2025.

[7] D. Squires and C. Anderson, “U.S. health care from a global perspective: Spending, outcomes, and social factors,” *The Commonwealth Fund*, 2015. [Online]. Available: https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#Health%20expenditures%20per%20capita,%20U.S.%20dollars,%20PPP%20adjusted,%202022. Accessed: Feb. 27, 2025.

[8] S. A. Lavarreda, E. R. Brown, and C. D. Bolduc, “Underinsurance in the United States: An interaction of costs to consumers, benefit design, and access to care,” *Annual Review of Public Health*, 2009. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev.publhealth.012809.103655. Accessed: Feb. 27, 2025.

[9] K. E. Thorpe, L. Allen, and P. Joski, “The role of chronic disease, obesity, and improved treatment and detection in accounting for the rise in healthcare spending between 1987 and 2011,” *Applied Health Economics and Health Policy*, vol. 15, no. 4, pp. 437–447, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s40258-015-0164-7. Accessed: Mar. 12, 2025.