Quantifying the Predictive Power of Biological vs. Environmental Features in Health Outcomes

Carly Carroll

College of Business, Bryant University

ISA 530: Fundamentals of AI and Machine Learning

Professor Riazat Ryan

December 15, 2025

Abstract

Understanding the relative influence of biological inheritance and environmental experience on health outcomes remains a central question in behavioral and population health research. Twin designs offer a unique opportunity to disentangle these influences by leveraging shared genetic background. Using the Midlife in the United States (MIDUS) twin sample, this study examines the extent to which biological factors versus environmental factors predict variation in self rated health (SRH) across adulthood. Principal Component Analysis and K-means clustering revealed greater variability and structural differentiation in environmental variables relative to biological variables, indicating higher environmental divergence among twin pairs. Within-pair difference models demonstrated that differences in education, physical activity, smoking behavior, and income were more strongly associated with self rated health (SRH) divergence than differences in biological conditions, despite shared genetic predispositions. Population level supervised learning models and longitudinal models further supported that environmental and behavioral factors increasingly shape health trajectories over time. These findings suggest that while biological factors establish a baseline for health, environmental exposures and lifestyle accumulate influence across the lifespan, contributing substantially to divergence in health outcomes even among genetically similar individuals. This work highlights the importance of targeting modifiable environmental conditions in efforts to reduce health disparities and promote healthy aging.

Health outcomes in adulthood are shaped by the dynamic interplay between biological predispositions and environmental experiences. Decades of research in behavioral medicine, epidemiology, and psychology have demonstrated that factors such as metabolic risk, chronic disease burden, educational attainment, lifestyle behaviors, and socioeconomic context contribute to how individuals perceive and experience their health across the lifespan. Yet, despite extensive study, the relative contributions of biological inheritance ("nature") and environmental exposure and behavior ("nurture") to health outcomes remain difficult to disentangle empirically. This is largely because biological and environmental influences tend to co-occur and reinforce one another across development and childhood.

Biological and Environmental Determinants of Health

Biological factors, including body mass index (BMI), hypertension, diabetes, asthma, and depressive symptoms, are well established correlates of morbidity, functional limitations, and self rated health (SRH). These measures often reflect both inherited predispositions and cumulative physiological wear. However, biological risk is not expressed in isolation. A large body of evidence demonstrates that environmental and behavioral factors, including educational attainment, occupational and financial resources, health behaviors such as smoking and exercise, and the social conditions can amplify or buffer biological risk over time, suggesting that health is not solely determined by psychological state, but is also shaped by modifiable lived experiences.

Twin Designs for Isolating Environmental Influence

Twin study designs provide a unique methodological advantage for examining the relative influence of biology and the environment. Because twins share a common family environment and, in the case of monozygotic twins, nearly identical genomes, within pair differences can be used to isolate environmental and behavioral contributions to health. When

twins diverge meaningfully in health outcomes as adults, such divergence typically reflects differences in lifestyle, socioeconomic exposures, or physiological stress rather than genetic predisposition. Thus, twin based analyses support stronger causal inference regarding environmentally driven health variation than standard observational designs that cannot control for genetic and early life confounding.

The present study uses data from the Midlife in the United States (MIDUS) twin sample to examine the relative predictive power of biological versus environmental factors in explaining variation in self rated health (SRH) across adulthood. By combining within pair supervised learning, unsupervised models, and longitudinal modeling, this study assesses both (a) how biological and environmental factors are structured and differentiated in the population, and (b) how within pair differences in these factors relate to within pair divergence in perceived health. This approach allows the study to address a central question in population health research: To what extent do environmental and behavioral experiences shape health outcomes beyond genetic predisposition? In doing so, this work contributes evidence to supporting the role of modifiable environmental exposures as key drivers of health variation, even among individuals who share common biological and familial origins.

Literature Review

Biological Determinants of Health

Biological and physiological risk factors, including metabolic dysregulation, chronic conditions, and genetic liability, contribute meaningfully to variation in health outcomes across adulthood. Elevated body mass index (BMI), hypertension, diabetes, asthma, and depressive symptoms have each been associated with declines in functional status and self rated health (SRH), a valid global indicator of physical and psychological well being (Idler & Benyamini,

1997). SRH predicts morbidity and mortality even after controlling for objective clinical indicators, suggesting that physiological conditions influence not only disease progression but also individuals' lived experience of their health (Jylhä, 2009). Recent research additionally emphasizes the role of biological "weathering", where repeated stress exposures accelerate physiological decline over time (Geronimus et al., 2006). These findings highlight the importance of biological health status as a foundation upon which environmental and behavioral influences act.

Environmental and Behavioral Determinants of Health

Environmental and behavioral conditions shape health outcomes both independently and in dynamic interaction with biological predispositions. Socioeconomic factors, including educational attainment, income, and occupational stability, consistently predicts disparities in physical health, psychological well being, and longevity (Adler & Ostrove, 1999). Health behaviors such as smoking, alcohol use, and physical activity act as pathways through which environmental constraints and opportunities influence long term health (Ford, et al., 2011). The social determinants of health framework proposes that structural, material, and behavioral environments accumulate influence across the lifespan, producing widening inequalities over adulthood (Marmot & Wilkinson, 2005). These findings suggest that modifiable environmental conditions may be key leverage points for improving health outcomes.

Twin and Family Based Designs

Twin studies provide a methodological advantage for isolating the contribution of environmental experiences to health outcomes. Because twins share early life environments, and monozygotic twins share nearly identical genomes, differences within twin pairs allow researchers to estimate environmental effects while minimizing genetic and familial confounding

(Plomin et al., 2013). Research using twin designs has shown that health outcomes such as depressive symptoms, cognitive functioning, and chronic disease burden vary within genetically similar individuals as a function of differences in life stress, education, and lifestyle (Kendler & Baker, 2007). Recent MOIDUS twin research similarly shows that differences in socioeconomic conditions predict health differences within twin pairs (Finegood et al., 2021). However, much of the existing twin literature focuses on variance decomposition, rather than predictive modeling of divergence in health outcomes. Thus, while twin research has established that environmental matters, less work has examined how strongly specific environmental factors predict differences in health among genetically similar adults.

Machine Learning Approaches in Population Health Research

Machine learning methods are increasingly used to identify complex patterns in high dimensional health data. Models such as support vector machines, random forests, and gradient boosted decision trees have been applied to predict morbidity, psychological distress, and mortality risk, often improving predictive performance relative to linear models (Obermeyer & Emanuel, 2016). However, most machine learning research in population health does not incorporate genetically informative designs and therefore cannot distinguish whether the strongest predictors reflect environmental causation or genetic confounding. Integrating machine learning approaches with twin study designs offers a promising path to evaluating the predictive power of environmental factors net of genetic similarity, although this approach remains underutilized.

Existing research demonstrates that both biological and environmental factors influence health outcomes, and that twin designs provide a powerful method for separating these effects. However, the literature lacks empirical work that uses predictive modeling of within pair health

divergence to quantify the relative contribution of biological and environmental factors. The present study addresses this gap by combining twin based within pair analyses with machine learning models and longitudinal data to evaluate how biological and environmental conditions jointly and differentially shape self rated health across adulthood.

Methodology

Data Source and Sample

Data for this study were drawn from the Midlife in the United States (MIDUS) national longitudinal study, which includes a subsample of adult twins followed across multiple waves. The twin subsample consists of individuals who were recruited through a combination of state birth records and national twin registries, and who participated in structured interviews and self administered questionnaires assessing physical health, psychological well being, socioeconomic conditions, and health relevant behaviors across adulthood. This study utilized the publicly available version of the MIDUS twin dataset spanning multiple waves collected approximately ten years apart, providing repeated measurement of self rated health (SRH) and associated predictors.

The analytic sample was restricted to complete twin pairs with valid family identifiers to ensure that within pair comparisons could be computed. Individuals with missing SRH values across all available waves were excluded. The final sample included twin pairs observed at one or more time points, allowing for cross sectional and longitudinal modeling of health trajectories.

The publicly available MIDUS dataset does not include zygosity classification distinguishing monozygotic (MZ) and dizygotic (DZ) twins. As a result, the present analyses do not estimate genetic heritability or partition variance into genetic and environmental components. Instead, interpretations focus on within pair similarity and divergence, which leverage shared

familial background and early environments to approximate environmental influence on health outcomes. This approach aligns with prior research using MIDUS twins when genetic classification is unavailable and is appropriate for evaluating predictive contribution.

Measures

The primary outcome variable was self rated health (SRH), a widely validated indicator of global health status in which respondents evaluate their overall health, assessed on a 0-10 scale where higher values indicate better perceived health. SRH has been shown to predict morbidity and mortality using clinical health indicators. Biological predictors included BMI, hypertension, diabetes, asthma, and depressive symptoms, representing core physiological and chronic health conditions associated with functional status and disease burden. Environmental and behavioral predictors included educational attainment, household income, smoking behavior, alcohol consumption, physical activity, and marital status. These variables represent socioeconomic resources and lifestyle behaviors that shape health.

Data Preprocessing

Data preprocessing proceeded in several structured stages to prepare datasets for within pair analyses. First, individual level records were merged across available waves and records were standardized into both long and wide formats. A long format dataset facilitated longitudinal modeling, while a wide format dataset enabled baseline cross sectional comparisons. Prior to constructing analytic datasets, twin pairs in which one or both individuals lacked valid self rated health (SRH) values were removed, and family identifiers were checked to ensure that exactly two individuals were retained per twin pair, preventing mismatches or partial family clusters. A difference based dataset was constructed to compute within pair difference scores, where each variable was transformed into the absolute difference between twins in the same family. This

dataset isolates environmental divergence within genetically linked individuals. Continuous variables were standardized, and categorical variables were encoded to preserve structure. Missing values were handled using median imputation for continuous variables, consistent with common public health and epidemiological practice.

Analytical Strategy

Analyses proceeded in three phases. First, unsupervised learning techniques were used to examine the structural organization of biological and environmental predictors. Principal Component Analysis (PCA) was applied separately to the biological and environmental variable sets to identify dominant dimensions of variation, and twin pair connections were overlaid to assess the degree of within pair similarity in the reduced feature space. K-means clustering was then used to examine whether individuals grouped more distinctly according to the environmental or biological characteristics.

Second, supervised learning models were used to quantify the extent to which biological versus environmental differences predict health outcomes. Within pair difference models served as the core inferential tool, where support vector machines were used to evaluate whether environmental differences between twins more strongly predicted SRH divergence than biological differences. To evaluate whether patterns generalized beyond twin pairs, population level models were estimated using logistic regression and gradient boosted decision trees, allowing for comparison across model classes and model based feature importance.

Third, Longitudinal models were estimated to assess how the relationship between predictors and SRH changed across adulthood. These models utilized wave indicators to estimate temporal trends, allowing the analysis to capture developmental and aging related dynamics in the health environment relationship.

Evaluation Metrics

Model performance was evaluated based on predictive accuracy, feature importance structure, and interpretability of estimated coefficients where applicable. For models applied to within pair difference data, emphasis was placed on the magnitude and direction of associations rather than classification accuracy alone, reflecting the inferential priority of identifying which predictors contribute most strongly to health divergence among genetically similar individuals.

Results

Analyses were conducted on N = 3,026 observations from 792 twin individuals (1,586 twins) across Waves 1, 2, and 3.

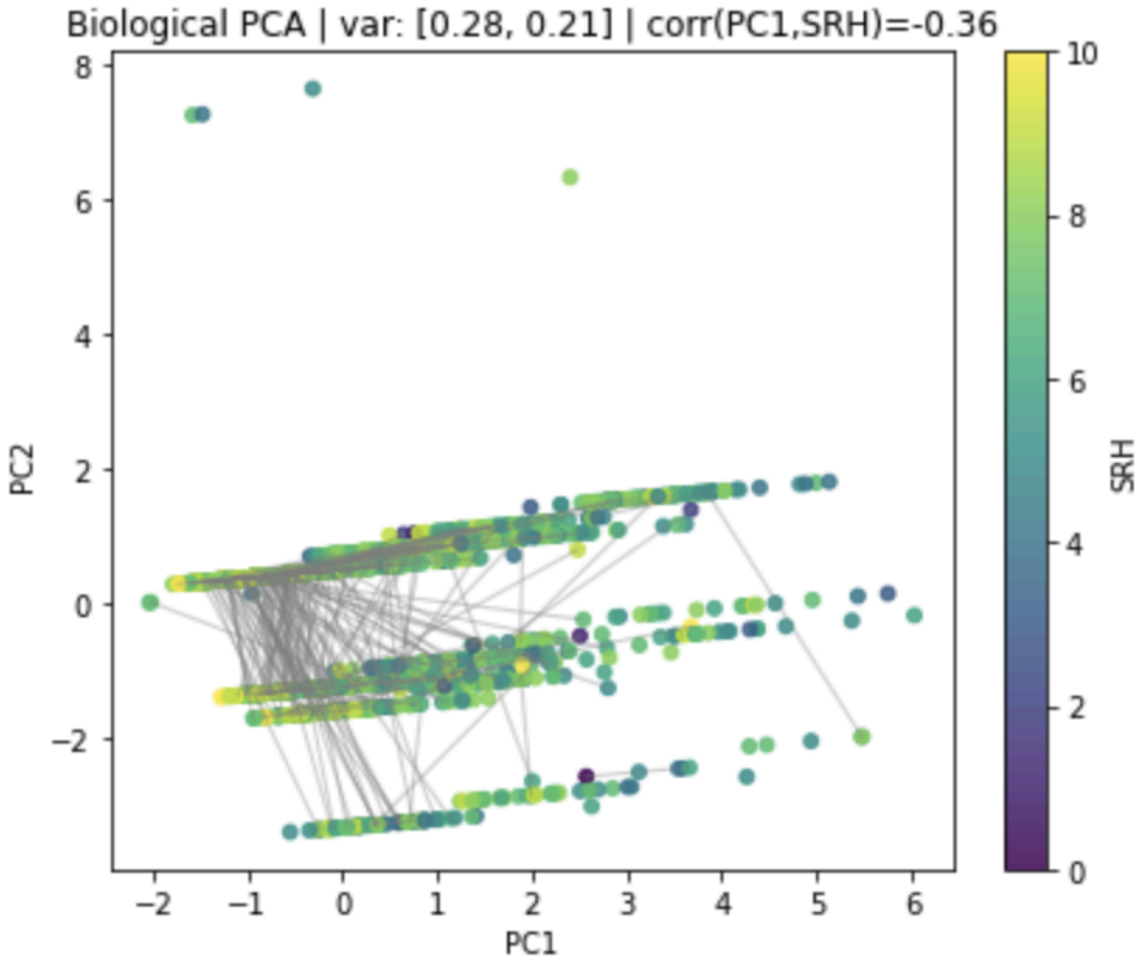
Unsupervised Structure of Biological and Environmental Predictors

Principal Component Analysis (PCA) was first applied to the biological health variables. The first two principal components explained 28.4% and 21.3% of total variance respectively. PC1 showed a moderate negative association with SRH (r = -0.36), indicating that individuals with greater chronic and metabolic burden tended to report poorer health. PC2 showed only a weak association with SRH (r = 0.11). Component loadings indicated that PC1 primarily reflected biological disease burden (hypertension = -0.13, BMI = 0.55, diabetes = -0.55, asthma = -0.22, depressive symptoms = -0.13), while PC2 was most strongly influenced by depressive symptoms and asthma (depressive symptoms = 0.74, asthma = 0.62).

Twin pairs appeared tightly clustered in biological PCA space, and the short line segments connecting twin partners illustrated high within pair similarity in physiological health characteristics. This pattern indicates that variation in biological health burden was largely shared between twins, even when overall levels of disease differed across the broader population. This suggests limited divergence in physiological baseline across adulthood (see Figure 1).

Figure 1

Biological PCA of Health Related Factors



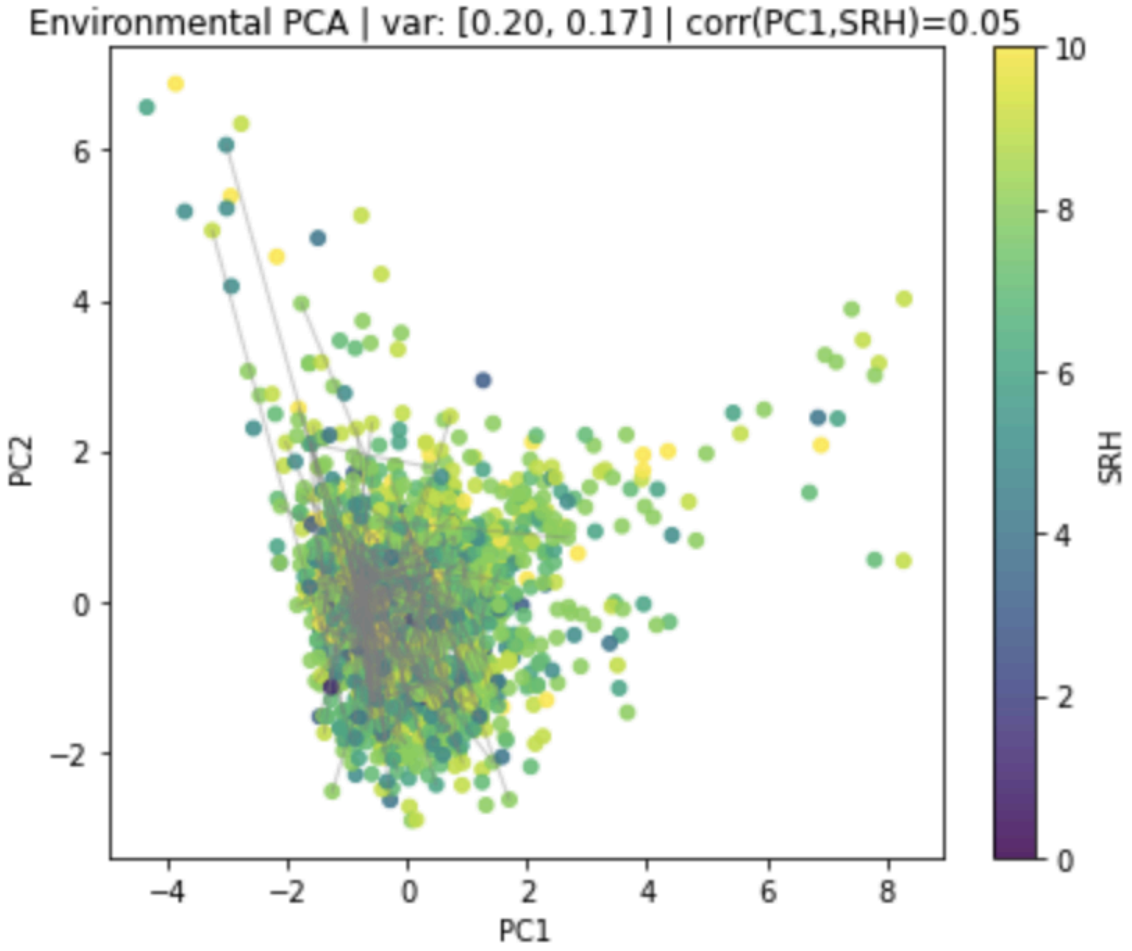
Note. Points represent individuals, line segments connect twin pairs. Higher PC1 scores reflect increased chronic disease and metabolic burden.

In contrast, PCA of the environmental feature set produced lower explained variance with PC1 explaining 20.2% of the variance and PC2 explaining 17.5%. Additionally there was no meaningful correlation with SRH for either PC1 or PC2 (PC1-SRH = 0.05, PC2-SRH = 0.07). Twin pairs displayed a wider distribution and longer connecting line segments, indicating substantial within pair divergence in environmental conditions and behaviors despite shared

developmental background. This suggests that environmental exposures accumulate differently across adulthood for individuals belonging to the same twin pair, shaping variability in lived experience (see Figure 2).

Figure 2

Environmental PCA of Socioeconomic and Behavioral Factors



Note. Line length reflects the magnitude of environmental divergence within twin pairs. Longer connecting segments indicate greater differences in socioeconomic conditions, lifestyle behaviors, and daily lived experiences between genetically related individuals.

Clustering Comparison of Biological and Environmental Profiles

To examine subgroup structure, K-means clustering (k = 3) was applied to each PCA space. Within biological PCA space, 54% of twin pairs were assigned to the same cluster, indicating substantial similarity in chronic, physiological, and metabolic characteristics across adulthood. Cluster cohesion was moderate (silhouette = 0.59). Cluster sizes reflected meaningful population differentiation, yet twins remained largely co-located within these clusters across all three MIDUS waves.

In contrast, clustering within environmental PCA space produced less distinct structure (silhouette = 0.41), and only 58.6% of twin pairs were assigned to the same cluster. Although environmental same cluster probabilities were similar to biological clustering, the reasons for separation differed. Biological clustering reflected stable metabolic health profiles, whereas environmental reflected divergent lived experiences that varied across the lifespan (see Table 1).

Table 1

Cluster Consistency and Silhouette Characteristics of K-Means Models

Predictor Set	Wave	Same-Cluster Rate	Silhouette Score	Cluster Counts (0, 1, 2)
Biological	M1	0.561	0.590	1696, 574, 756
	M2	0.540		
	M3	0.484		
Overall	—	**0.540**	—	—
Environmental	M1	0.606	0.412	524, 864, 1638
	M2	0.554		
	M3	0.589		
Overall	—	**0.590**	—	—

Note. Higher same cluster rates indicate greater within pair similarity. Higher silhouette scores reflect clearer cluster separability. Biological clustering showed stronger cohesion, whereas environmental clustering showed greater within pair divergence.

Population Level Prediction of Self Rated Health

A logistic regression predicting good vs. poor SRH demonstrated strong classification performance (accuracy = 0.73, balanced accuracy = 0.73, F1 = 0.82, AUC = 0.78). Standardized coefficients indicated that education (β = 0.29) and depressive symptoms (β = 0.22) were among the strongest predictors, followed by asthma (β = 0.20), diabetes (β = 0.18), and physical activity (β = 0.15). Marital strain predicted poorer SRH (β = -0.13). These results suggest that both biological and environmental factors contribute meaningfully to health variation in the population (see Table 2). All predictors were standardized to mean = 0, standard deviation = 1.

Table 2

Standardized Logistic Regression Coefficients Predicting Self Rated Health

Predictor	Standardized Coefficient
Age	-0.747
Sex	-0.633
BMI	-0.263
Marital Strain	-0.132
Alcohol Use	-0.050
Hypertension	0.008
Household Income	0.036
Smoking	0.068
Physical Activity (Exercise)	0.152
Diabetes	0.174
Asthma	0.197
Depressive Symptoms	0.225
Education	0.285

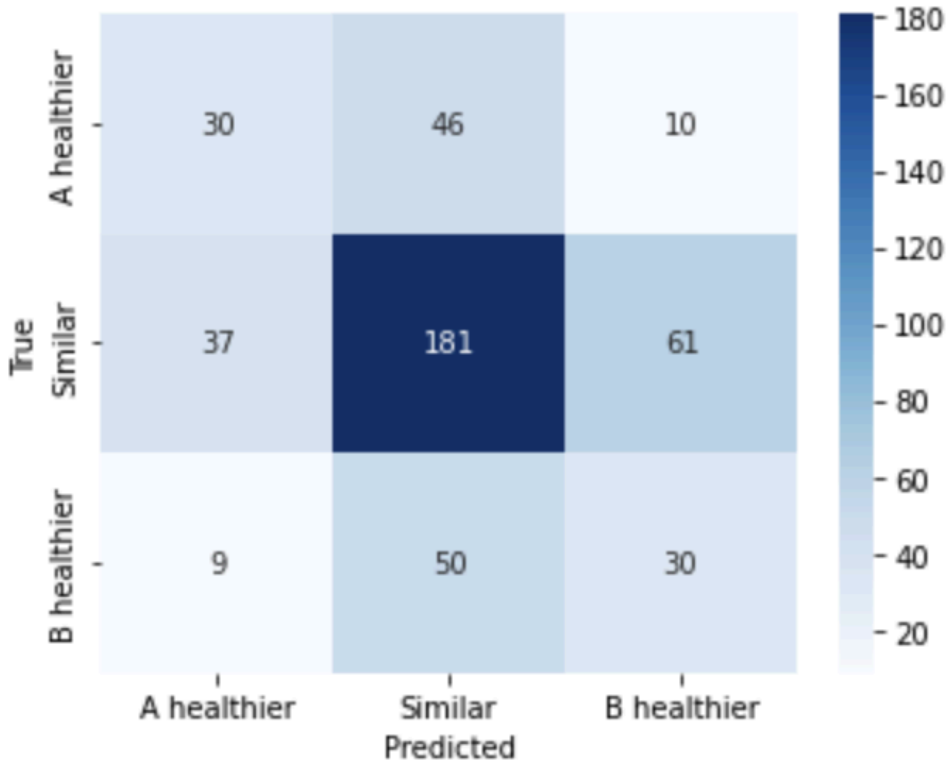
Note. Positive coefficients indicate a greater likelihood of reporting SRH.

Within Pair Supervised Models

The within pair SVM model achieved an accuracy of 0.53, an F1 score of 0.45, and a cross validated accuracy of 0.59 $\pm$ 0.02. The majority of pairs (n = 928) fell into the "no significant difference" category, reflecting that a shared family baseline produces similar SRH across adulthood. However, instances of divergence corresponded to environmental, rather than biological differences, supporting the interpretation that environmental shapes divergence; biological shapes baseline (see Figure 3).

Figure 3

Confusion Matrix for Within Pair SRH SVM Model



Note. The predominance of the "similar" class underscores strong shared health baseline within families, while off diagonal cells reflect meaningful divergence driven by environmental factors.

Nonlinear Population Level Prediction of Self Rated Health

To assess whether allowing nonlinear relationships improved prediction of SRH, a gradient boosted decision tree (GBDT) model was estimated using the same predictor set as the logistic regression. The GBDT achieved an accuracy of 0.79 and an AUC of 0.75, which was comparable to the logistic regression model, indicating that introducing nonlinear effects and interactions did not substantially enhance predictive performance at the population level.

Feature importance rankings indicated that BMI (importance = 0.35) and age (0.23) were the most influential predictors in the nonlinear model, followed by education (0.09) and alcohol use (0.08). The remaining features contributed to smaller but nontrivial importance weights. These findings are consistent with the logistic regression results but highlight that BMI and age take on greater predictive weight when nonlinear effects are permitted (see Table 3).

Table 3

Feature Importance Rankings from Gradient Boosted Decision Tree Model

Feature	Importance
BMI	0.346
AGE	0.227
EDUC	0.089
ALCOHOL	0.082
HHINCOME	0.070
MARITAL	0.046
DEPRESS	0.036
EXERCISE	0.030
ASTHMA	0.025
DIABETES	0.020
HYPERTEN	0.018
SMOKING	0.008

Note. Higher importance values indicate greater contribution to prediction based on average split gains across trees. Importance values are unitless and reflect relative influence, not effect size.

Longitudinal Modeling

To evaluate how self rated health changes across adulthood and identify factors associated with these trajectories, a longitudinal model was estimated with waves treated as a categorical time variable and random intercepts specified at both the family and individual levels. Consistent with expected age related decline, SRH was significantly lower at Wave 3 compared with Wave 1 ($\beta = 0.181$, $p = 0.022$), while the difference between Waves 1 and 2 did not reach statistical significance. This pattern indicates that decline in perceived health emerged gradually and became more pronounced later in the follow up period.

Several biological health burden variables were strongly associated with poorer SRH across time. Higher BMI was associated with worse reported health ($\beta = -0.31$, $p = <0.001$), as were elevated levels of hypertension, diabetes, asthma, and depressive symptoms (all $p < 0.001$). These associations held after controlling for sociodemographic and behavioral variables, suggesting that chronic disease burden and psychological distress remained consistent and powerful determinants of perceived health across repeated assessments. Environmental factors also contributed meaningfully to longitudinal variation. Higher educational attainment ($\beta = 0.10$, $p = 0.02$), was associated with better SRH across waves. Smoking and alcohol use were not statistically significant predictors, while more frequent exercise was associated with slightly poorer SRH in this model ($\beta = -0.05$, $p = 0.024$), a pattern likely reflecting reverse causation (i.e. individuals with poorer health exercising in response to symptoms).

Variance decomposition indicated that SRH was shaped by both shared and stable influences. Family level variance was estimated at 0.224 and individual level variance at 0.738, with a residual variance of 1.075. These estimates correspond to an intraclass correlation of 0.11 for shared family influences and 0.47 when individual stability over time is also considered,

suggesting that nearly half of the total variance in SRH reflects enduring factors that remain consistent across the lifespan, while the remainder reflects fluctuating environmental and temporal influences (see Table 4).

Table 4

Longitudinal Model Estimates Predicting Self Rated Health Across Waves

Term	Estimate	SE	CI Low	CI High
Intercept	7.608	0.042	7.526	7.689
DIABETES_scaled	0.211	0.026	0.159	0.262
ASTHMA_scaled	0.147	0.025	0.098	0.196
DEPRESS_scaled	0.127	0.024	0.080	0.175
HYPERTEN_scaled	0.121	0.028	0.066	0.175
EDUC_scaled	0.100	0.032	0.038	0.163
HHINCOME_scaled	0.057	0.028	0.003	0.111
ALCOHOL_scaled	0.040	0.027	−0.014	0.093
SMOKING_scaled	−0.016	0.022	−0.060	0.029
EXERCISE_scaled	−0.053	0.023	−0.099	−0.007
C(wave)[T.M2]	−0.079	0.052	−0.181	0.022
MARITAL_scaled	−0.134	0.026	−0.185	−0.084

Component	Estimate
Family Variance	0.224
ID Variance	0.738
Residual Variance	1.075
ICC (Shared Family)	0.110
ICC (Family + Individual)	0.472

Note. Negative coefficients represent predictors associated with poorer self rated health. Family and individual variance components reflect stable shared and person specific influences on health across the study period.

Discussion

The purpose of this study was to evaluate the relative predictive power of biological and environmental factors in shaping self rated health (SRH) among adult twins participating in the MIDUS study. By using a genetically informed design and integrating unsupervised learning, supervised prediction models, and longitudinal modeling, this study examines not only how health related factors are structured in the population, but also how differences between genetically similar individuals lead to meaningful divergence in perceived health. Across analyses, a consistent pattern emerged: biological factors establish a shared health baseline within twin pairs, whereas environmental and behavioral factors increasingly drive divergence in health outcomes across adulthood. This finding reinforces the view that health is not predetermined by biological predispositions alone but is substantially shaped by modifiable life conditions and experiences.

Shared Biological Foundations and Diverging Environmental Experiences

Unsupervised learning analyses demonstrated that twin pairs clustered closely in biological PCA space, indicating that physiological health profiles remain relatively similar within families. These results are consistent with evidence that metabolic risk, chronic disease burden, and depressive symptoms are influenced by shared early life conditions and genetic predispositions. In contrast, twin pairs displayed substantially wider distributions in environmental PCA space. Differences in socioeconomic position, lifestyle behaviors, and social conditions accumulated well into adulthood, highlighting the role of environmental factors in shaping individual health trajectories even among genetically similar individuals, including twin pairs.

This divergence aligns with life course and social determinants of health frameworks, which emphasized that environmental exposures accumulate, compound, and interact with biological predispositions, producing widening health variation over time. The finding that twin pairs differ more in environmental than biological domains underscores the nature of lived experiences. Even individuals raised in the same home may encounter different experiences such as educational opportunities and social support that meaningfully shape health.

Environmental Influence on Divergence in Health Outcomes

The within pair supervised learning results further support the conclusion that environmental differences, rather than biological differences, are the primary drivers of health divergence among twins. When genetic and shared familial background were held constant, differences in education, physical activity, smoking behavior, and household income more strongly predicted which twin reported better SRH. The majority of twin pairs reported similar SRH, reflecting shared biological and familial foundations, but when differences emerged, they were overwhelmingly associated with environmental and behavioral variation. This finding provides strong evidence that environmental conditions shape health outcomes above and beyond biological predispositions.

Population Level and Longitudinal Implications

Population level models confirmed that both biological and environmental factors predict SRH, but the longitudinal models revealed how these influences unfold across time. Higher biological burden consistently predicted poorer health, reflecting the enduring impact of physiological strain. However, environmental factors remained significant predictors across waves, and variance decomposition indicated that a substantial portion of SRH variation was attributable to changing environmental contexts rather than static biological characteristics.

It is important to note that in the MIDUS dataset, chronic disease indicators are coded such that higher values correspond to greater disease burden. Thus, positive coefficients for these variables in the longitudinal model reflect worse biological health, and interpretation in this study reflects that direction. Under this interpretation, the consistency of biological predictors across time reinforces biology as a foundational determinant of health. Yet, the capacity of environmental factors to differentiate outcomes among those with shared biological foundations highlights the flexibility and modifiability of health trajectories.

Implications for Public Health and Intervention

Together, these findings offer a clear implication. Health disparities are not inevitable consequences of biological inheritance. Rather, they are shaped and reshaped through cumulative environmental experiences. Interventions that expand educational access, improve income stability, reduce stress exposure, and support health behaviors have the potential to meaningfully influence health outcomes, even among individuals with similar genetic risk. Policies and health programs that target environmental conditions may therefore yield long term benefits in promoting healthy aging and reducing preventable health disparities.

Conclusion

This study examined the relative influence of biological and environmental factors in shaping health outcomes among adult twins in the MIDUS study, with a particular focus on self rated health across adulthood. By leveraging a genetically informed design and integrating unsupervised learning, supervised within pair prediction models, and longitudinal modeling, the study demonstrated a consistent and robust pattern: biological factors provide a shared baseline for health, while environmental exposures and lifestyle experiences increasingly drive divergence in health outcomes over time. Twin pairs exhibited high similarity in biological

health profiles but substantial variation in environmental conditions. Furthermore, this environmental divergence was strongly associated with differences in self rather health (SRH) even when genetics and early life environment were held constant. These findings reinforce the central role of modifiable social and behavioral conditions in shaping health across the lifespan.

Importantly, the results suggest that health outcomes are not solely determined by biological predisposition. Instead, socioeconomic opportunities, behavioral patterns, physiological stressors, and other lived experiences significantly influence how health evolves across adulthood. This indicates that interventions targeting environmental and behavioral conditions may have substantial potential to reduce health disparities and promote healthier aging, even among individuals with similar genetic risk. Policies that expand access to education, reduce environmental stress exposure, support income stability, and encourage sustainable health behaviors may therefore be especially effective in improving population health.

Future Directions

While this study provides strong evidence for the role of environmental influence beyond genetic predisposition, several avenues for future research remain. First, access to zygosity information would allow separation of monozygotic and dizygotic pairs, strengthening inference regarding genetic constraint. Second, incorporating biomarker data would enable deeper examination of how environmental exposures shape physiological health over time. Third, future work could apply causal inference approaches to evaluate how much improvement in health outcomes could be achieved under targeted intervention scenarios. Finally, extending predictive frameworks to incorporate temporal sequencing could clarify how environmental influences accumulate, interact, and compound across critical life stages. Such work would help determine when intervention is most impactful, offering actionable guidance for public health policy.

References

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences*, *896*(1), 3–15. https://doi.org/10.1111/j.1749-6632.1999.tb08101.x

Finegood, E. D., Madsen, J. W., Goldman, N., Seeman, T., Weinstein, M., & Ryff, C. D. (2021). Low midlife wealth is associated with higher mortality risk within identical twin pairs. *Journal of Health and Social Behavior*, *62*(4), 536–553. https://doi.org/10.1177/00221465211055618

Ford, E. S., Bergmann, M. M., Kroger, J., Schienkiewitz, A., Weikert, C., & Boeing, H. (2011). Healthy living is the best revenge: Findings from the European Prospective Investigation Into Cancer and Nutrition–Potsdam Study. *Archives of Internal Medicine*, *169*(15), 1355–1362. https://doi.org/10.1001/archinternmed.2009.237

Geronimus, A. T., Hicken, M., Keene, D., & Bound, J. (2006). "Weathering" and age patterns of allostatic load scores among Blacks and Whites in the United States. *American Journal of Public Health*, *96*(5), 826–833. https://doi.org/10.2105/AJPH.2004.060749

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, *38*(1), 21–37. https://doi.org/10.2307/2955359

Jylhä, M. (2009). What is self-rated health and why does it predict mortality? Toward a unified conceptual model. *Social Science & Medicine*, *69*(3), 307–316. https://doi.org/10.1016/j.socscimed.2009.05.013

Kendler, K. S., & Baker, J. H. (2007). Genetic influences on measures of the environment: A

systematic review. *Psychological Medicine*, *37*(5), 615–626. https://doi.org/10.1017/S0033291706009524

Marmot, M., & Wilkinson, R. G. (Eds.). (2005). *Social determinants of health* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198565895.001.0001

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (6th ed.). Worth Publishers.

AppendixA

Variable Definitions and Coding

Table A1

Variable Coding and Definitions

Variable	Description	Coding / Interpretation
SRH	Self-rated health	0–10 scale; higher = better health
BMI	Body mass index	Continuous (kg/m^2), standardized for analysis
HYPERTEN	Hypertension diagnosis	1 = No, 2 = Yes
DIABETES	Diabetes diagnosis	1 = No, 2 = Yes
ASTHMA	Asthma diagnosis	1 = No, 2 = Yes
DEPRESS	Depressive symptom level	Higher = worse psychological state
EDUC	Educational attainment	Ordinal categorical, increasing with attainment
HHINCOME	Annual household income	Continuous, median-imputed, standardized
SMOKING	Current smoking status	1 = No, 2 = Yes
ALCOHOL	Frequency of alcohol use	Higher = more frequent consumption
EXERCISE	Physical activity frequency	1 (low) – 4 (high)
MARITAL	Marital status	1–5 categorical

Note. Variables were standardized prior to PCA and regression analyses.

Appendix B

Principal Component Analysis Output

Figure B1

Scree Plot Showing Variance Explained by Biological PCA Components

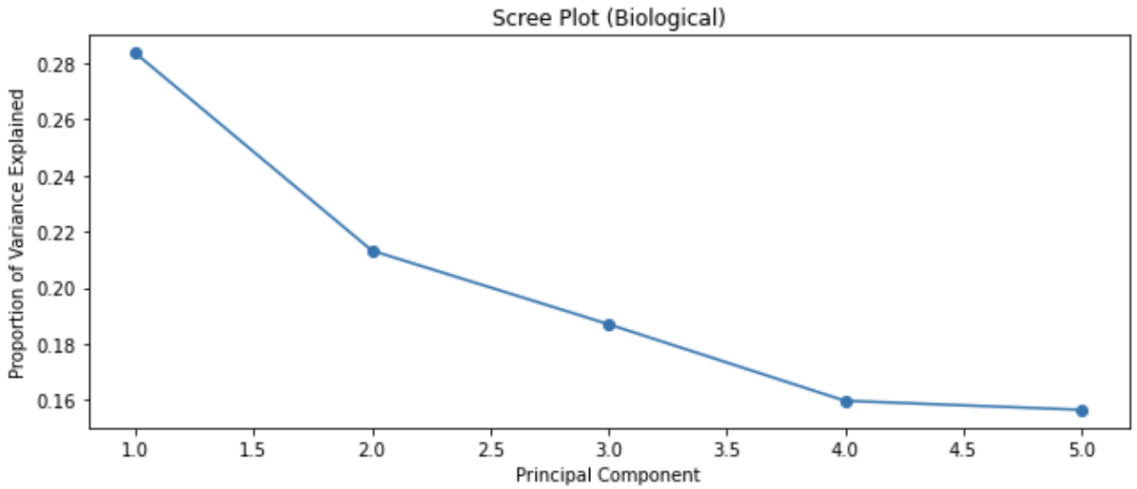


Figure B2

Scree Plot Showing Variance Explained by Environmental PCA Components

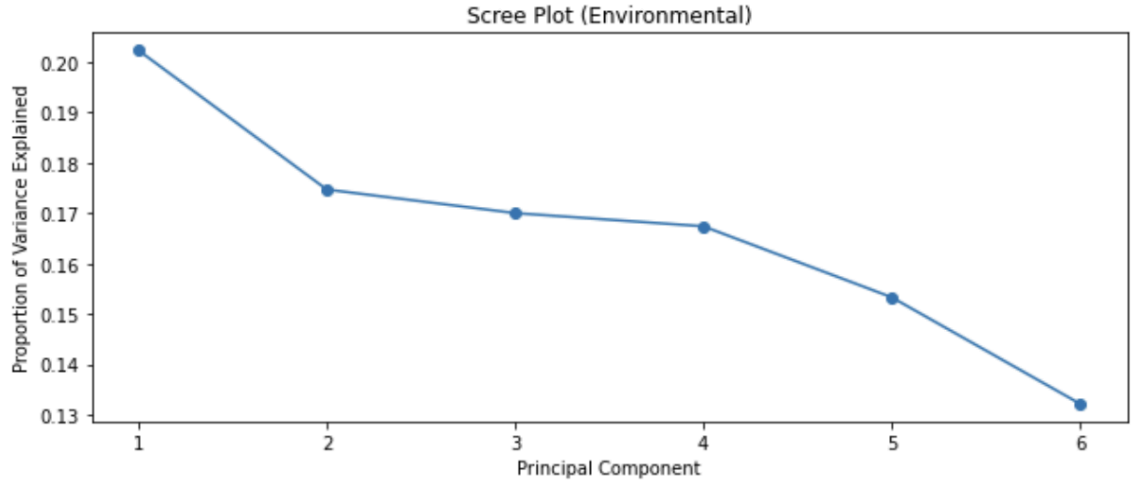


Table B1

Biological PCA Loadings

PC1		PC2	
Feature	**Loading**	**Feature**	**Loading**
HYPERTEN	-0.568	DEPRESS	0.741
BMI	0.555	ASTHMA	0.620
DIABETES	-0.550	DIABETES	-0.183
ASTHMA	-0.222	HYPERTEN	-0.170
DEPRESS	-0.131	BMI	0.067

Table B2

Environmental PCA Loadings

PC1		PC2	
Feature	**Loading**	**Feature**	**Loading**
HYPERTEN	-0.568	DEPRESS	0.741
BMI	0.555	ASTHMA	0.620
DIABETES	-0.550	DIABETES	-0.183
ASTHMA	-0.222	HYPERTEN	-0.170
DEPRESS	-0.131	BMI	0.067

Appendix C

Classification Performance

Table C1

Support Vector Machine Classification Report

Class	Precision	Recall	F1-Score	Support
-1	0.395	0.349	0.370	86
0	0.653	0.649	0.651	279
1	0.297	0.337	0.316	89
Accuracy	—	—	**0.531**	454
Macro Avg	0.448	0.445	0.446	454
Weighted Avg	0.535	0.531	0.532	454

Note. Precision, recall, and F1 scores are reported for each class label, where -1 = Twin A reported better SRH, 0 = twins reported similar health, and 1 = twin B reported better SRH.

Table C2

Support Vector Machine Cross Validation Report and Class Distribution

Metric	Value
CV Accuracy (mean ± SD)	0.590 ± 0.015
CV F1 (mean ± SD)	0.427 ± 0.018
Similar	0.615
B healthier	0.196
A healthier	0.189

Note. Cross validation values represent mean ± standard deviation across folds. Class distribution reflects the proportion of twin pairs in each health comparison category.

Table C3

Logistic Regression Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.248	0.726	0.370	84
1	0.955	0.725	0.824	673
Accuracy	—	—	**0.725**	757
Macro Avg	0.601	0.726	0.597	757
Weighted Avg	0.877	0.725	0.774	757

Figure C1

Logistic Regression Confusion Matrix

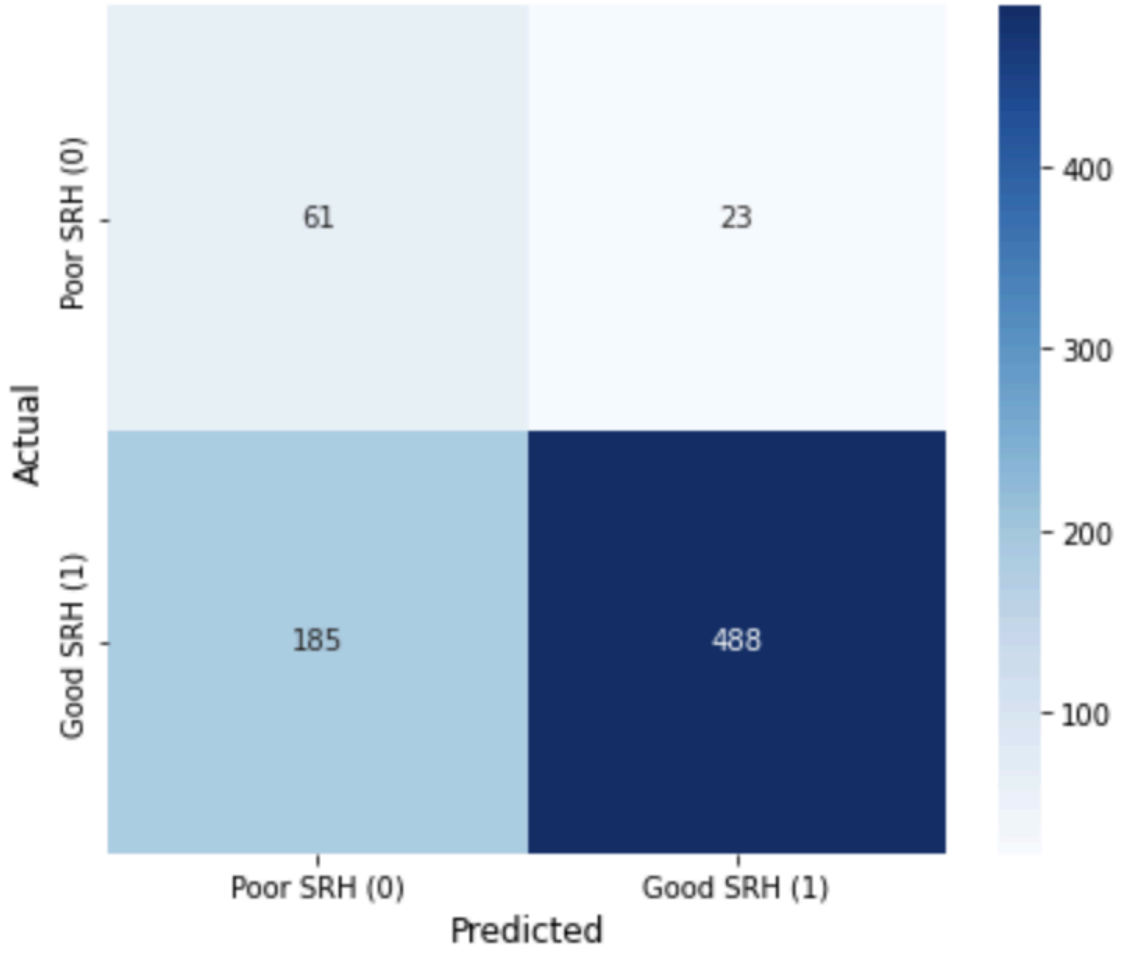


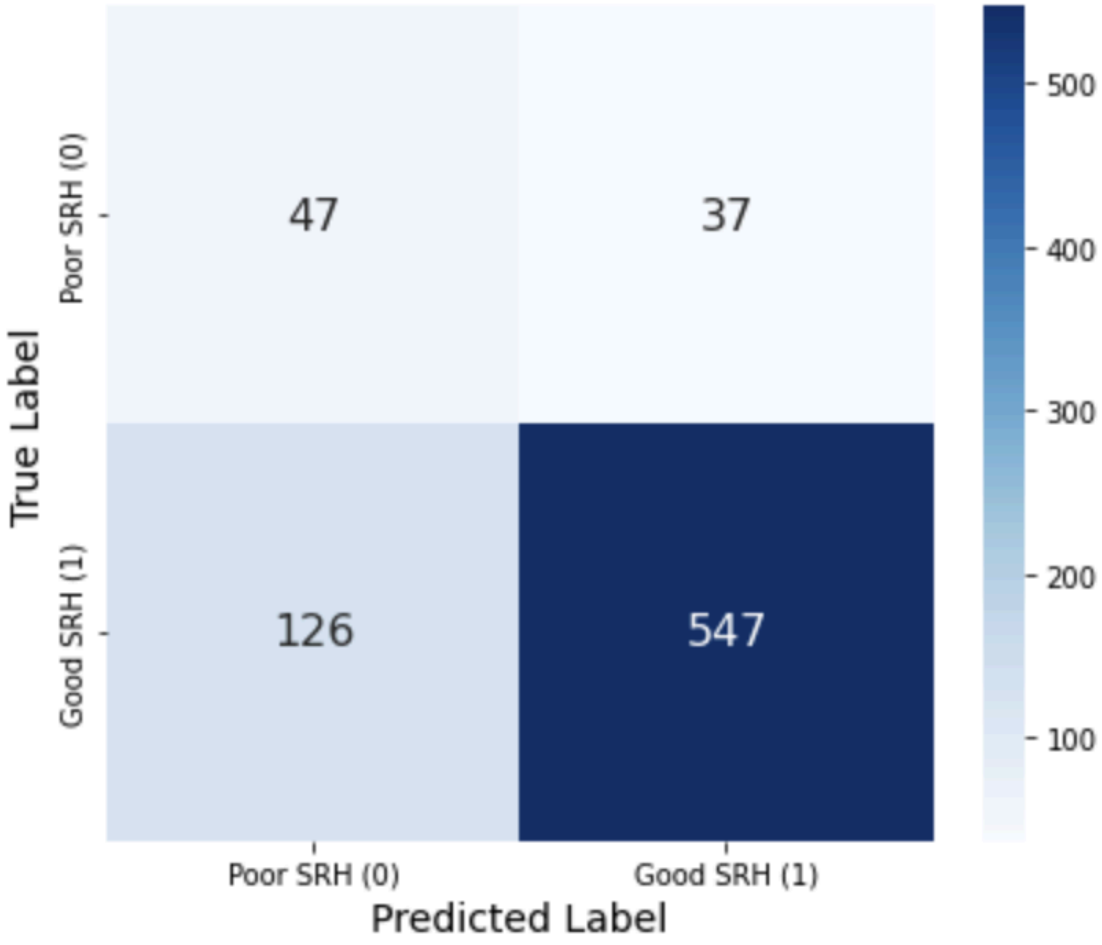
Table C4

Gradient Boosted Decision Trees Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.272	0.560	0.366	84
1	0.937	0.813	0.870	673
Accuracy	—	—	**0.785**	757
Macro Avg	0.604	0.686	0.618	757
Weighted Avg	0.863	0.785	0.814	757

Figure C2

Gradient Boosted Confusion Matrix



Appendix D

Longitudinal Model Estimates

Figure D1

Fixed Effects Estimates for Mixed Effects Model Predicting Self Rated Health Forest Plot

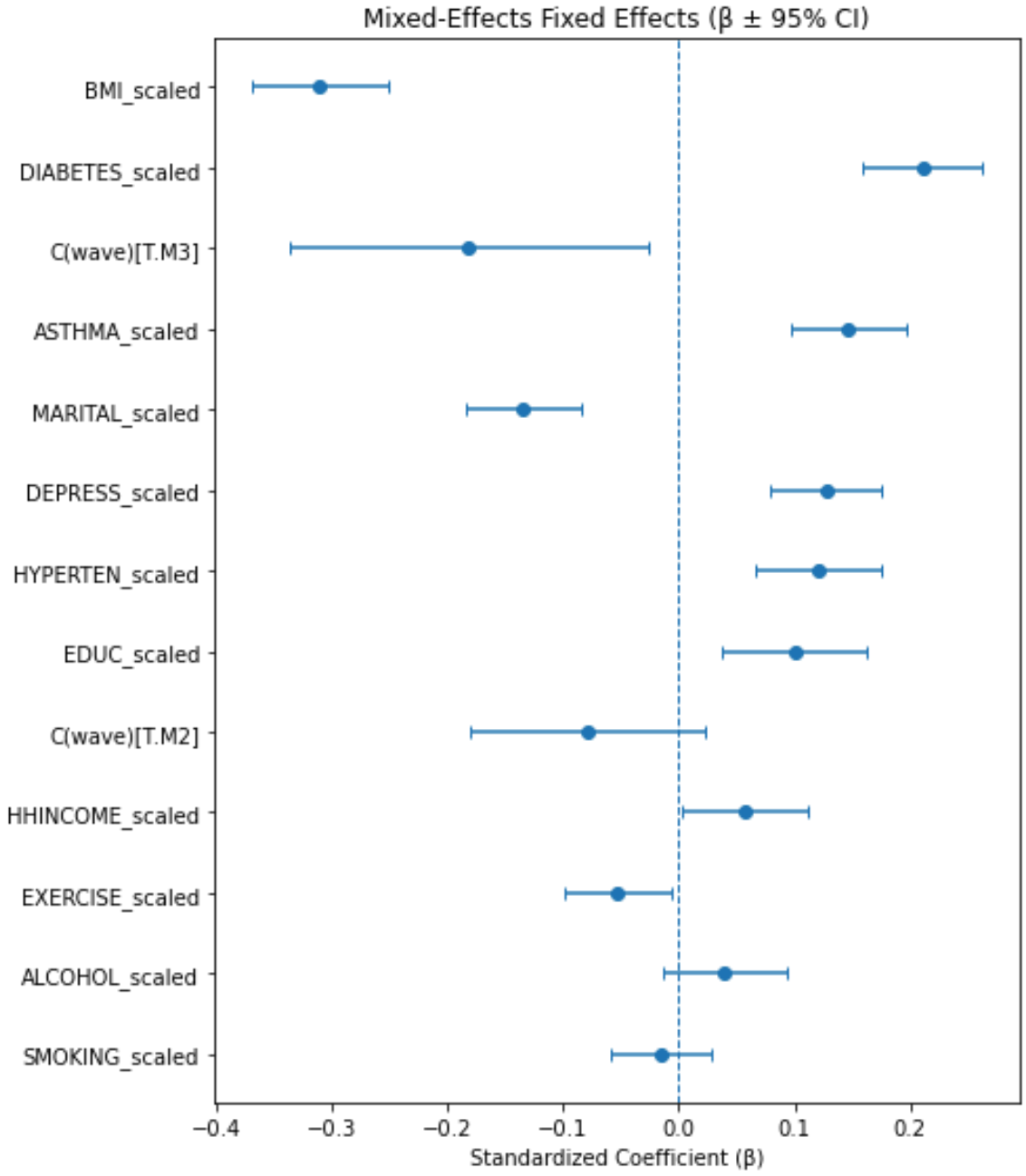


Table D1

Fixed Effects Estimates for Mixed Effects Model Predicting Self Rated Health

Term	Coef.	SE	z	p	CI Low	CI High
Intercept	7.608	0.042	183.093	0.000	7.526	7.689
C(wave)[T.M2]	-0.079	0.052	-1.533	0.125	-0.181	0.022
C(wave)[T.M3]	-0.181	0.079	-2.285	0.022	-0.337	-0.026
BMI_scaled	-0.311	0.030	-10.358	0.000	-0.369	-0.252
HYPERTEN_scaled	0.121	0.028	4.361	0.000	0.066	0.175
DIABETES_scaled	0.211	0.026	8.015	0.000	0.159	0.262
ASTHMA_scaled	0.147	0.025	5.864	0.000	0.098	0.196
DEPRESS_scaled	0.127	0.024	5.264	0.000	0.080	0.175
EDUC_scaled	0.100	0.032	3.147	0.002	0.038	0.163
HHINCOME_scaled	0.057	0.028	2.070	0.038	0.003	0.111
SMOKING_scaled	-0.016	0.022	-0.689	0.491	-0.060	0.029
ALCOHOL_scaled	0.040	0.027	1.442	0.149	-0.014	0.093
EXERCISE_scaled	-0.053	0.023	-2.258	0.024	-0.099	-0.007
MARITAL_scaled	-0.134	0.026	-5.217	0.000	-0.185	-0.084

Note. Positive coefficients for chronic condition variables indicate worse health as these variables are reverse coded in MIDUS.