Healthcare Data Analysis Using MEPS 2020

Data-Driven Insights into Health Expenditure, Chronic Disease, and Mental Health

AMS 597 – Statistical Computing

Instructor: Dr. Silvia Sharna

Stony Brook University

Junsung Rhee	116740814
Keun Young Yoon	116517818
Dhananjay Sharma	116740841
Parth Pidadi	116237374

Contents

0.1 Introduction

Healthcare has long been a critical domain in both social and economic discussions. In the United States, this issue is particularly complex due to the role of private insurance systems, which determine access to essential care for a large portion of the population. Rising healthcare costs and gaps in insurance coverage have led to increasing concerns about equity and affordability. This prompted our team to ask: **Can data-driven insights help reduce healthcare expenditures and, in doing so, broaden access to insurance coverage?**

To explore this question, we utilized data from the **Medical Expenditure Panel Survey (MEPS)**, specifically the **2020 Full-Year Consolidated Data File (HC-224)** of the Household Component. MEPS is jointly administered by the *Agency for Healthcare Research and Quality (AHRQ)* and the *National Center for Health Statistics (NCHS)*. It is one of the most comprehensive and reliable sources of information on healthcare utilization, expenditure, insurance coverage, and individual-level demographic and behavioral characteristics in the United States. The 2020 file consolidates responses from three overlapping panels across multiple interview rounds, providing a nationally representative snapshot of healthcare dynamics over a full calendar year.

The analysis began by examining which components of medical expenditure most significantly contribute to total healthcare spending. Office-based doctor visits and prescription drug expenditures constitute the largest components of total healthcare spending. Inpatient hospitalizations, while less frequent, contribute significantly to overall costs due to their high per-case expense. This led us to a critical turning point in our study: **Could these high-cost components be reduced through earlier intervention?**

With this in mind, we turned our attention to two well-established drivers of long-term healthcare burden: chronic disease and psychological distress. Chronic diseases, if left unmanaged, often lead to costly hospitalizations and long-term care. Psychological distress, similarly, is linked to increased medical utilization and poorer overall health outcomes. We hypothesized that early identification of individuals at risk—based on demographic and behavioral patterns—could help reduce future costs and improve population health.

To guide our analysis, we focused on the following three research questions:

- **Research Question 1:** Identifying which components of medical expenditure most significantly contribute to individuals' total annual healthcare spending.
- **Research Question 2:** Examining the association between demographic factors and chronic disease prevalence.

- **Research Question 3:** Exploring the relationship between lifestyle factors, health behaviors, and psychological distress through multiple analytical approaches.

Through this multi-stage, data-driven approach, our goal is to derive actionable insights that inform preventive strategies, promote efficient resource use, and ultimately support more equitable access to care.

0.2 Dataset Overview

0.2.1 Source of the Dataset

This project utilizes the **MEPS HC-224: 2020 Full-Year Consolidated Data File**, a public-use dataset released by the *Agency for Healthcare Research and Quality (AHRQ)*. The Medical Expenditure Panel Survey (MEPS) is a nationally representative survey that provides detailed information on healthcare utilization, expenditures, insurance coverage, chronic conditions, behavioral factors, and demographics for the U.S. civilian noninstitutionalized population.

The 2020 MEPS dataset was chosen for its comprehensive scope and rich individual-level information, enabling in-depth analysis of how health status, demographic traits, and health behaviors interact with healthcare costs and access. Its wide coverage of both clinical and behavioral variables makes it particularly suitable for our multi-dimensional approach, which aims to uncover predictors of health spending, chronic disease, and psychological distress.

0.2.2 Structure and Scope

The original dataset contains **27,805 individuals (rows)** and **1,451 variables (columns)**. For the purposes of this study, we selected **38 variables** most relevant to our research goals. All 27,805 individuals were retained to preserve statistical representativeness. The selected variables fall into five main categories and are detailed in Table 1.

Table 1: Selected Variables from MEPS HC-224 Dataset with Descriptions and Types

Variable Name	**Description**	**Type**
DUPERSID	Person ID (DUID + PID)	Numerical
FCSZ1231	Family Size Responding 12/31 CPS Family	Numerical

Continued on next page

Table 1 – continued from previous page

Variable Name	**Description**	**Type**
REGION20	Census Region as of 12/31/20	Categorical
AGE20X	Age as of 12/31/20 (Edited/Imputed)	Numerical
DOBYY	Date of Birth: Year	Numerical
SEX	Sex	Categorical
RACEV2X	Race (Edited/Imputed)	Categorical
MARRY20X	Marital Status-12/31/20 (Edited/Imputed)	Categorical
EDUCYR	Years of Educ When First Entered MEPS	Numerical
HIBPDX	High Blood Pressure Diagnosis (>17)	Categorical
CHDDX	Coronary Heart Disease Diagnosis (>17)	Categorical
MIDX	Heart Attack (MI) Diagnosis (>17)	Categorical
STRKDX	Stroke Diagnosis (>17)	Categorical
CHBRON53	Chronic Bronchitis Last 12 Months (>17)-R5/3	Categorical
CHOLDX	High Cholesterol Diagnosis (>17)	Categorical
CANCERDX	Cancer Diagnosis (>17)	Categorical
DIABDX_M18	Diabetes Diagnosis	Categorical
JTPAIN53_M18	Joint Pain Last 12 Months (>17) - RD 5/3	Categorical
ARTHDX	Arthritis Diagnosis (>17)	Categorical
NOSMOK42	Dr Advise Smoking in Home is Bad (0-17) - R4/2	Categorical
PHYEXE53	Mod/Vigorous Physical Exercise 5X Week (>17) - RD 5/3	Categorical
OFTSMK53	How Often Smoke Cigarettes (>17) - RD 5/3	Categorical
K6SUM42	SAQ 30 Days: Overall Rating of Feelings	Categorical
ADSLEEP42	How Often Trouble With Sleep	Categorical
ADKALC42	SAQ 12 Months: Asked Alcohol Consumption	Categorical
ADNUMDRK42	SAQ 12 Months: Number of Drinks on Typical Day	Categorical
FAMINC20	Family's Total Income	Numerical
POVLEV20	Family Income as Percent of Poverty Line	Numerical
INSCOV20	Health Insurance Coverage Indicator 2020	Categorical
INSURC20	Full Year Insurance Coverage Status 2020	Categorical
MCARE20X	Covered by Medicare - 12/31/20 (Edited)	Numerical
TOTTCH20	Total Direct Charges (Excl. Prescribed Medicines)	Numerical
TOTEXP20	Total Health Care Expenditures in 2020	Numerical

Continued on next page

Table 1 – continued from previous page

Variable Name	**Description**	**Type**
OBDEXP20	Office-Based Doctor Expenditures	Numerical
OPTEXP20	Office-Based Therapist Expenditures	Numerical
ERDEXP20	ER Doctor Expenditures	Numerical
RXEXP20	Prescribed Medicines Expenditures	Numerical
IPDEXP20	Total Hospital Stay Doctor Expenditures 2020	Numerical

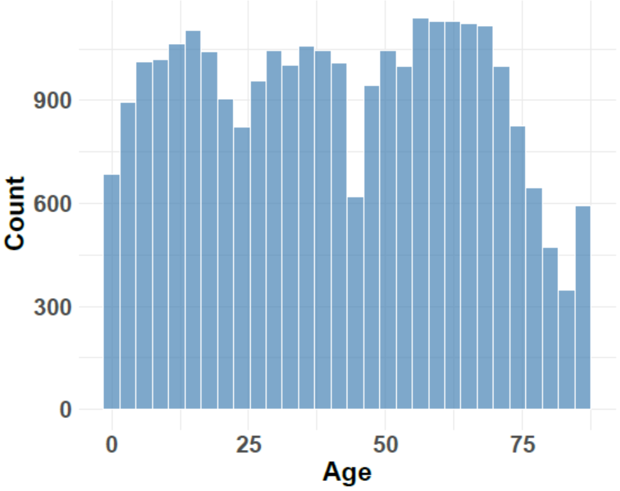


Figure 1: Age Distribution

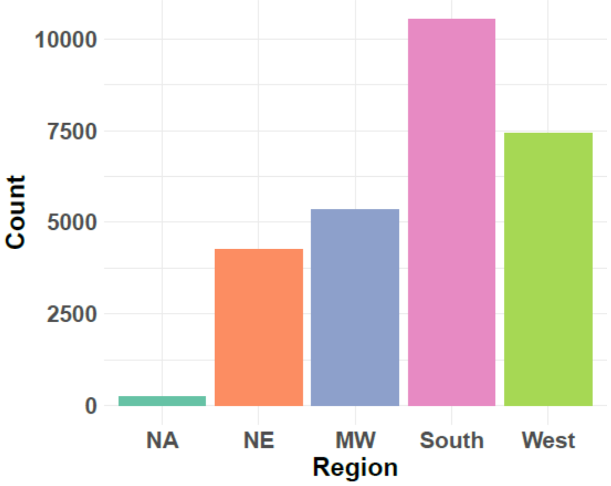


Figure 2: Regional Distribution

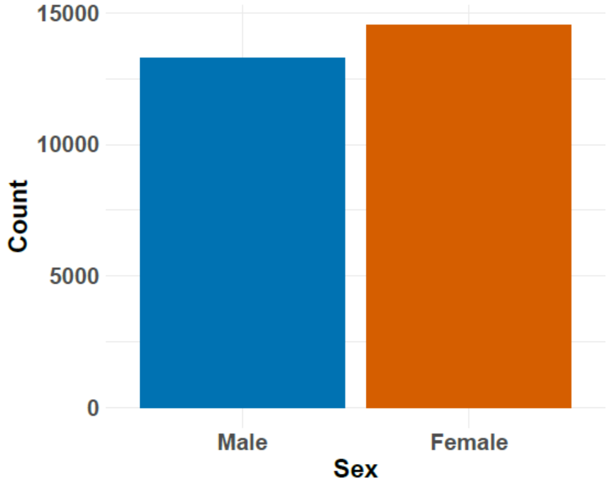


Figure 3: Sex Distribution

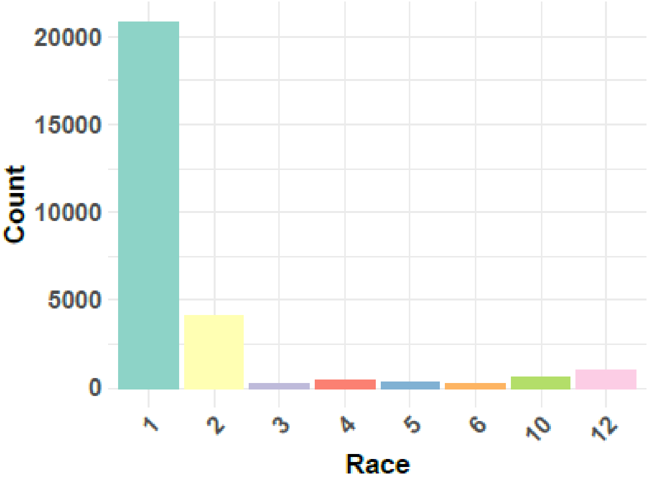


Figure 4: Race Distribution

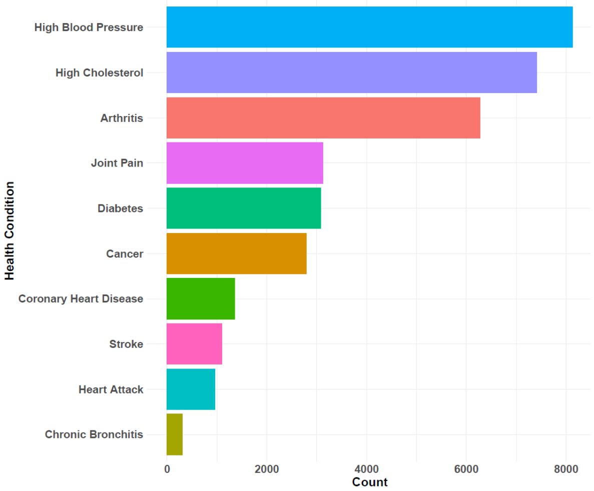


Figure 5: Chronic Disease Prevalence

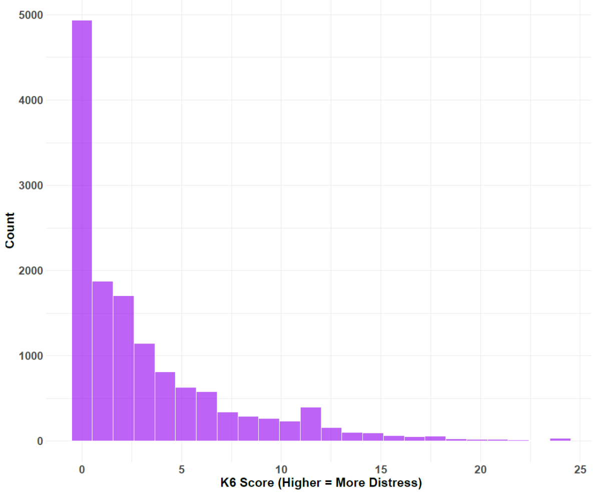


Figure 6: Psychological Distress (K6 Score)

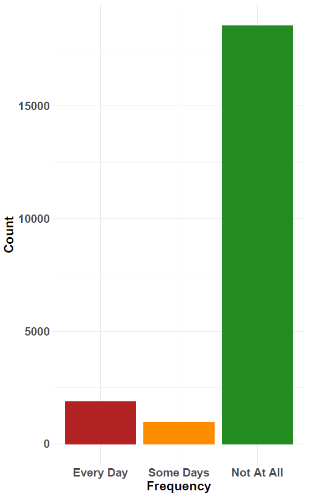


Figure 7: Smoking Frequency

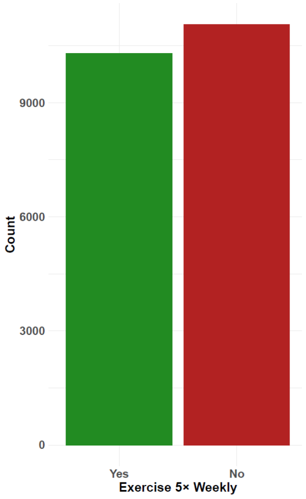


Figure 8: Exercise Frequency (5X/Week)

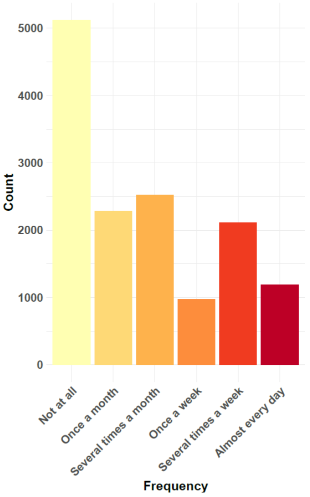


Figure 9: Frequency of Sleep Issues

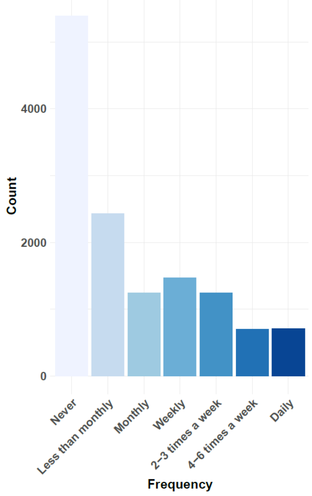


Figure 10: Alcohol Consumption Frequency

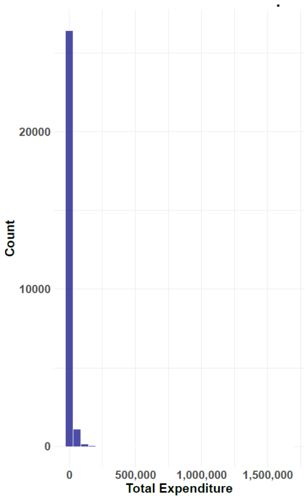


Figure 11: Total Healthcare Expenditure

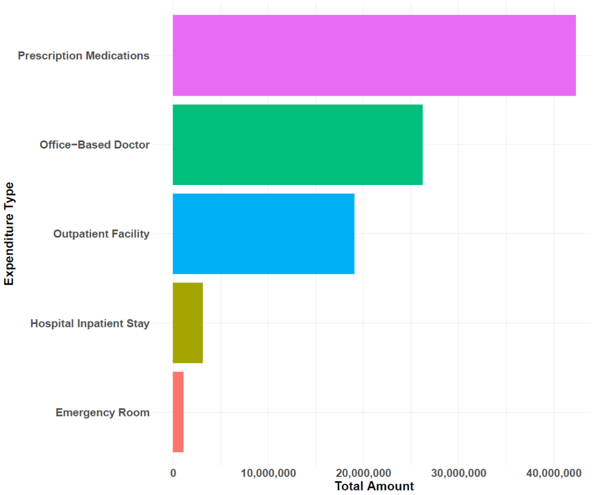


Figure 12: Healthcare Expenditure by Type

0.3 Methodology and Implementation

This section details the comprehensive methodological approach taken to answer each of the three research questions. The analysis was implemented in R using various statistical, machine learning, and visualization tools. Each research question involved distinct data wrangling, modeling strategies, and interpretation pipelines. Below, we describe the full implementation pipeline for each question in detail.

0.3.1 RQ1: Expenditure Analysis

Objective The objective of this analysis is to identify which components of medical expenditures most significantly drive an individual's total annual healthcare costs. By evaluating the relative contributions of key categories—such as inpatient services, emergency care, and prescription drugs—the goal is to inform targeted cost-containment strategies and improve the allocation of healthcare resources.

1.1 Variable Selection and Preprocessing The dependent variable in this analysis is TOTEXP20, representing each individual's total annual healthcare expenditure. Five major spending categories were selected as predictors:

- OBDEXP20: Office-Based Doctor Expenditures
- OPTEXP20: Office-Based Therapist Expenditures
- ERDEXP20: Emergency Room Doctor Expenditures
- IPDEXP20: Inpatient Hospital Doctor Expenditures

- RXEXP20: Prescription Drug Expenditures

All six variables—the dependent variable and five predictors—were log-transformed using the formula log(x + 1) to reduce right-skewness and stabilize variance. Observations with zero values for TOTEXP20 were excluded, not only because they are incompatible with log transformation, but also because the data distribution contains a high proportion of zero-expenditure cases. To enable a more accurate assessment of spending patterns and cost drivers, the analysis focused on individuals who had incurred at least some healthcare costs. The final analytic sample includes individuals with non-zero total expenditures and complete data across all selected categories.

Table 2: Summary Statistics of Healthcare Expenditure Components

Variable	Mean	Median	Min	Max	Q1	Q3	SD	Zeros	Zero%
TOTEXP20	7655.66	1761	1	1662894	495	6052.75	24004.18	0	0.00
OBDEXP20	1143.88	223	0	696525	0	765.00	9010.19	6608	28.84
OPTEXP20	829.44	0	0	365193	0	16.00	5783.59	17161	74.89
ERDEXP20	46.44	0	0	9952	0	0.00	272.16	20487	89.40
IPDEXP20	136.25	0	0	49056	0	0.00	997.71	21441	93.56
RXEXP20	1844.70	70	0	1642905	0	576.25	13809.37	7198	31.41

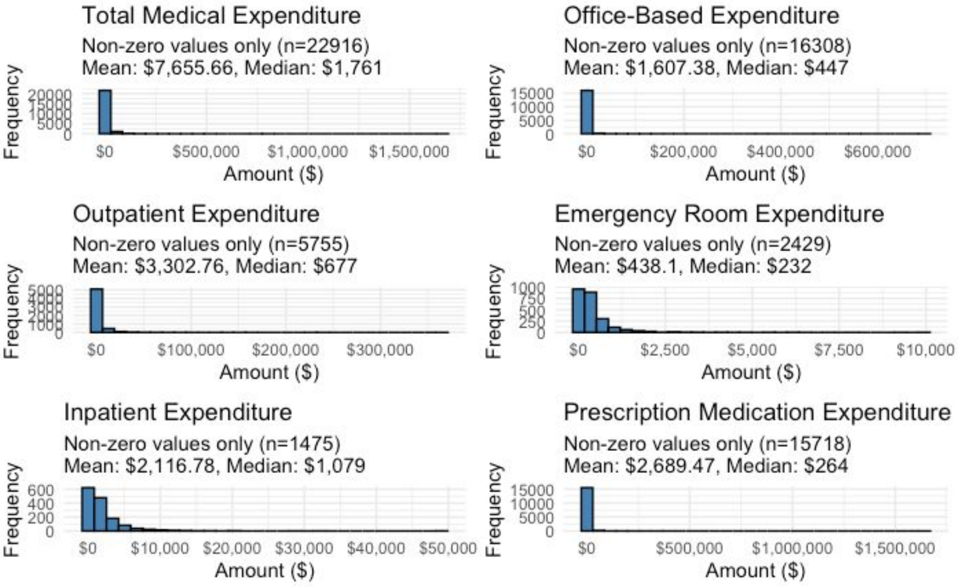


Figure 13: Distribution of Medical Expenditures by Category

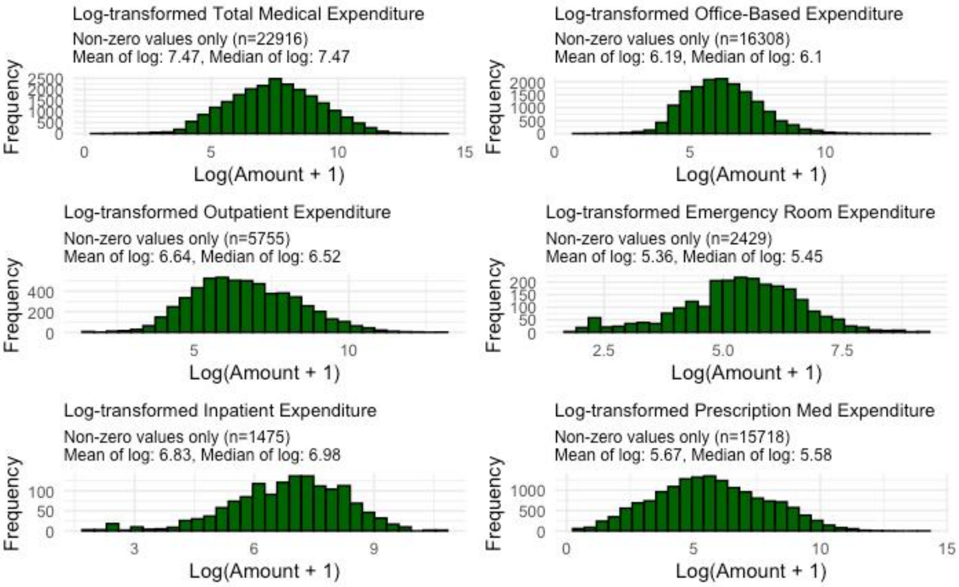


Figure 14: Distribution of Log-Transformed Medical Expenditures by Category

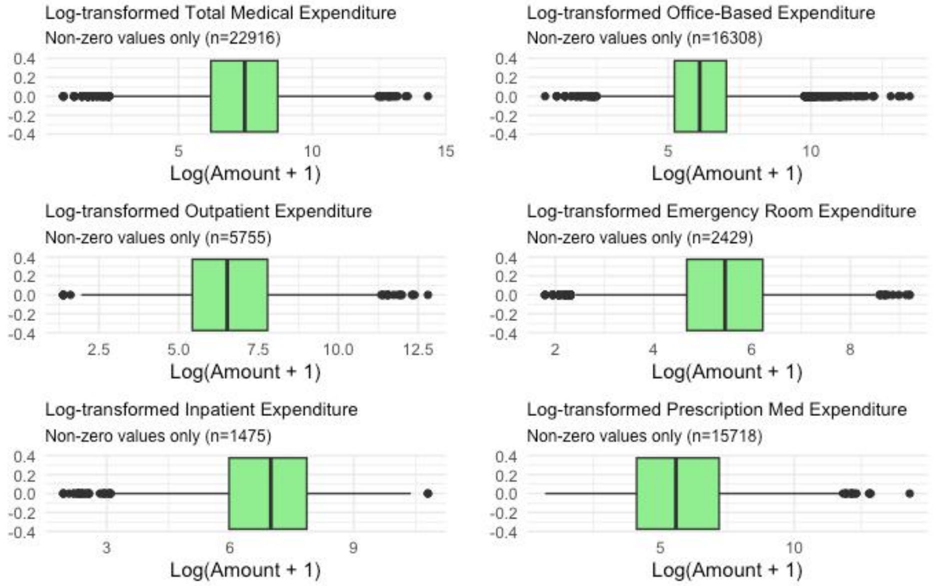


Figure 15: Boxplots of Log-Transformed Medical Expenditures by Category

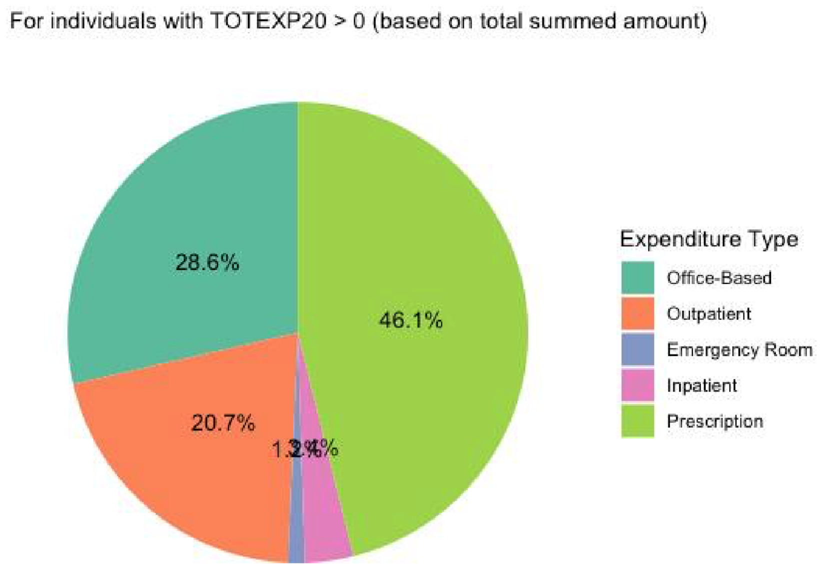


Figure 16: Total Expenditure Share by Category (Summed Across All Individuals)

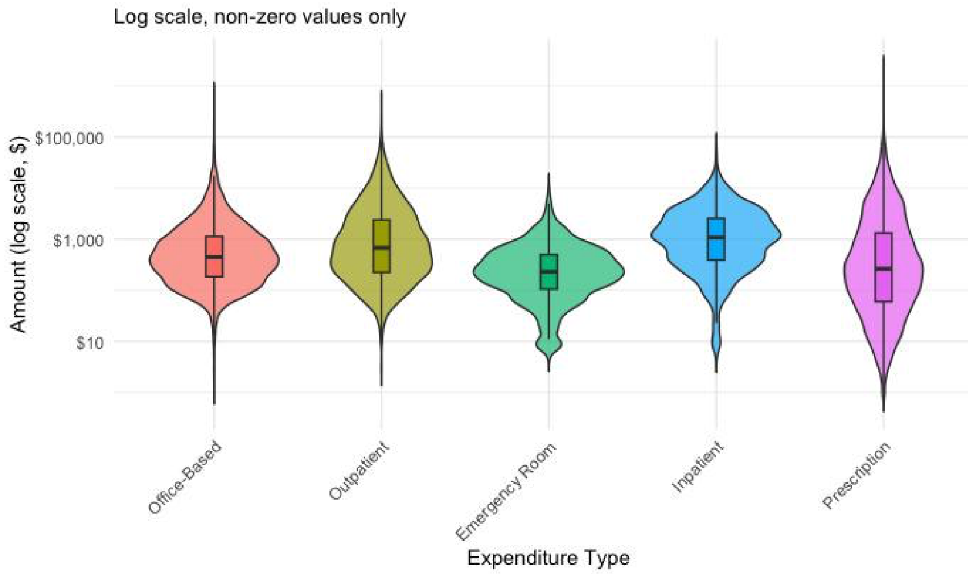


Figure 17: Violin Plot of Log-Scaled Medical Expenditures by Category

1.2 Analytical Approach A multiple linear regression model was estimated using the log-transformed total expenditure as the dependent variable and the five log-transformed cost components as predictors.

To examine the proportion of variance explained by each component, an analysis of variance (ANOVA) was performed on the fitted model. Multicollinearity among predictors was assessed by calculating Variance Inflation Factors (VIFs), all of which were found to be low, indicating minimal collinearity.

1.3 Predictive Modeling To assess predictive performance, two supervised regression models were developed:

- A multiple linear regression model using the log-transformed spending components
- An XGBoost regression model, a tree-based ensemble method capable of capturing non-linear relationships and interactions

The dataset was randomly split into training (70%) and test (30%) subsets. Both models were trained on the training data and evaluated on the test set.

Model performance was assessed using the following metrics:

- Root Mean Squared Error (RMSE)
- R-squared (R^2)

For the XGBoost model, feature importance metrics were extracted and visualized to identify the most influential spending categories contributing to total healthcare costs.

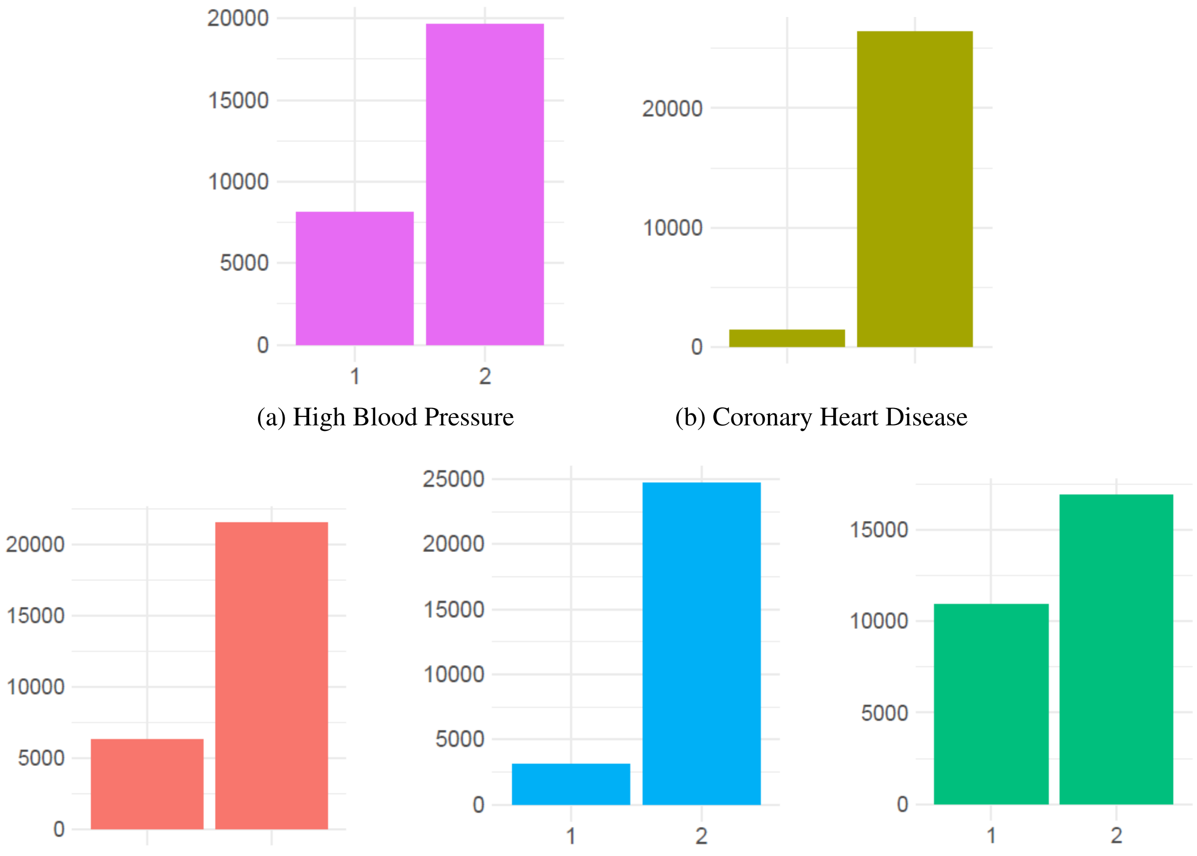
0.3.2 RQ2: Chronic Disease and Demographics

Objective The objective of this analysis is to examine whether demographic characteristics such as age, sex, race, geographic region, and family size are associated with the presence of chronic diseases. Identifying demographic risk factors for chronic illness can help inform more equitable and preventive public health strategies.

2.1 Variable Selection and Preprocessing The dependent variable for this analysis is **CHRONIC**, a binary indicator representing whether an individual had at least one of the following chronic conditions:

- HIBPDX: High Blood Pressure
- CHDDX: Coronary Heart Disease
- ARTHDX: Arthritis
- DIABDX_M18: Diabetes

Each condition was originally coded as 1 (Yes), 2 (No), and values less than 0 for Not Applicable or missing. All values less than 0 were recoded to 2 (No), based on the assumption that most of these correspond to ineligible younger individuals. A new variable **CHRONIC** was created and assigned a value of 1 if any of the four conditions was marked Yes, and 2 otherwise.



(a) High Blood Pressure (b) Coronary Heart Disease

(c) Arthritis (d) Diabetes (e) Chronic Conditions

Figure 18: Chronic Disease Distributions by Condition

The independent variables used in the model were:

- FCSZ1231: Family size
- REGION20: Census region, Northeast(1), Midwest(2), South(3), West(4)
- AGE20X: Age
- SEX: Sex
- RACEV2X: Race

Missing values were imputed using domain-informed rules: family size values of -1 were recoded to 1 (single-person household), region was imputed with the mode (Region 3), and age with the mean of all valid values. Categorical variables were converted to factors as needed.

2.2 Analytical Approach A logistic regression model was estimated using the full dataset (N = 27,805) to evaluate associations between the selected demographic variables and the likelihood of having a chronic disease. This approach enabled odds-based interpretation of each predictor.

2.3 Predictive Modeling To assess the predictive power of demographic variables, two supervised classification models were developed:

- A logistic regression model, which estimates the log-odds of chronic disease presence based on demographic predictors
- A random forest classifier, an ensemble learning method that captures complex, non-linear interactions and provides variable importance rankings

The data were randomly split into training (70%) and testing (30%) subsets. Both models were trained to predict the **CHRONIC** variable and evaluated on the held-out test set.

Model performance was assessed using the following metrics:

- Accuracy
- Sensitivity
- Specificity

For the random forest model, feature importance was extracted and visualized to identify the most influential predictors of chronic disease.

0.3.3 RQ3: Lifestyle and Distress

Objective The objective of this analysis is to investigate the relationship between lifestyle behaviors and psychological distress. Specifically, we aim to determine which behavioral factors are most associated with mental distress and to evaluate whether predictive models can accurately identify individuals at risk.

3.1 Variable Selection and Preprocessing The dependent variable is **K6SUM42**, representing an individual's psychological distress level, based on the Kessler-6 scale. For binary classification, individuals with scores greater than or equal to 13 were labeled as experiencing high distress (coded as 1); all others were coded as 0.

The independent variables consist of six behavioral indicators:

- PHYEXE53: Physical activity frequency
- ADSLEEP42: Sleep disturbance frequency

- OFTSMK53: Frequency of cigarette smoking
- ADKALC42: Alcohol consumption frequency
- NOSMOK42: Doctor's advice against smoking
- ADNUMDRK42: Number of alcoholic drinks consumed on a typical day

Negative response codes (e.g., -1, -7, -9) indicating refusals, non-responses, or inapplicable answers were recoded or imputed using reasonable defaults such as zero or modal values. Observations with missing outcome values were removed. The final dataset retained 27,805 individuals with complete data.

3.2 Analytical Approach Two main modeling strategies were applied to explore the association between lifestyle behaviors and psychological distress:

- **Principal Component Analysis (PCA):** PCA was performed on six behavioral variables to reduce dimensionality and uncover latent structures. The number of components retained was selected based on the cumulative proportion of explained variance. The resulting component scores were used as input features in downstream predictive modeling.
- **K-means Clustering:** K-means clustering was applied to the PCA component scores to segment individuals into behavioral pattern groups. The optimal number of clusters was determined using the elbow method, and each respondent was assigned to a cluster. These cluster labels were included as categorical predictors in one of the classification models.

3.3 Predictive Modeling To evaluate the predictive power of behavioral indicators, four classification models were developed to predict high psychological distress:

- Logistic regression using the original six behavioral variables
- Random forest classifier using the same behavioral variables
- Logistic regression using the principal component scores
- Logistic regression using cluster labels as a single categorical predictor

The dataset was split into training (70%) and testing (30%) subsets. All models were trained on the training data and evaluated on the test data to assess generalization performance.

Model performance was assessed using the following metrics:

- Accuracy
- Sensitivity
- Specificity
- Area Under the ROC Curve (AUC)

The ROC curves were generated for each model to visualize classification trade-offs across different thresholds. These visualizations are presented in the results section.

0.4 Results and Interpretation

This section presents the key findings from each of the three research questions, along with visual and statistical evidence. We discuss the observed trends, model performance metrics, and the real-world implications of each result. All results were derived from rigorous modeling, visualization, and interpretation in R.

RQ 1: Expenditure

Summary of Findings

- Among the five expenditure categories, office-based doctor visits (OBDEXP20) contributed the most to the total variance in healthcare spending, as indicated by ANOVA results. It also ranked second in XGBoost feature importance. However, its regression coefficient was relatively small, suggesting a frequent but low-cost nature of this service.
- Inpatient hospital expenditures (IPDEXP20) showed the largest regression coefficient, despite lower variance contribution and lower feature importance in XGBoost. This indicates that although less frequent, inpatient care incurs extremely high costs per episode, making it a critical cost driver.
- Prescription drug expenditures (RXEXP20) consistently ranked among the top in all three analyses—ANOVA, regression coefficients, and XGBoost importance—highlighting its strong influence in both frequency and unit cost. It is identified as a key driver of total medical spending.
- Emergency room (ERDEXP20) and therapist-based outpatient expenditures (OPTEXP20) had moderate to weak contributions across all methods and were not identified as primary cost drivers.

Term	Estimate
log_obd	0.1537
log_opt	0.1305
log_er	0.1266
log_ip	0.2303
log_rx	0.2202

Table 3: Coefficients

Term	SS
log_obd	18955.2
log_opt	7946.1
log_er	4180.4
log_ip	3783.8
log_rx	8666.3

Table 4: ANOVA (SS)

Term	VIF
log_obd	1.224
log_opt	1.112
log_er	1.176
log_ip	1.180
log_rx	1.302

Table 5: VIF

Figure 19: Linear Regression on Full Sample: Coefficients, ANOVA, and Multicollinearity Diagnostics

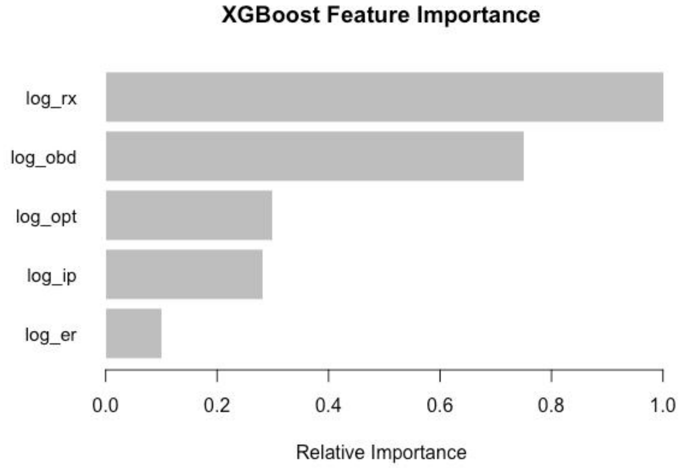


Figure 20: XGBoost Feature Importance for Chronic Disease Prediction

Metric	Linear Regression	XGBoost
RMSE	1.136	0.949
R^2	0.591	0.715

Table 6: Comparison of Regression Model Performance Metrics

Interpretation The combination of regression modeling, ANOVA, and XGBoost results provides a multidimensional view of what drives total healthcare costs. While prescription drugs emerged as a consistent and powerful driver due to their high cost and usage, inpatient care posed the greatest cost burden on a per-case basis despite lower frequency. On the other hand, office-based doctor visits, although individually inexpensive, represent a major source of total variance due to their high utilization.

These findings suggest that policy and insurance design should prioritize both high-frequency and high-cost categories—particularly prescription medications for chronic conditions and catastrophic hospitalization events—while recognizing that high-usage, low-cost services like routine doctor visits still exert a significant cumulative impact.

RQ 2: Chronic Disease

Summary of Findings

- A logistic regression model fitted on the full dataset identified **age** and **family size** as the most influential demographic predictors of chronic disease. Age showed a strong positive association with disease presence, while household size showed a negative relationship. Several racial and regional differences were also statistically significant.
- For prediction purposes, both logistic regression and random forest models were developed. Logistic regression demonstrated slightly higher **sensitivity** (ability to identify people with chronic disease), while random forest achieved better **specificity** (ability to detect those without chronic disease). Both models yielded similar overall accuracy around 82%.
- Variable importance analysis from the random forest model reaffirmed that **age** and **family size** were the dominant predictors, consistent with the logistic regression findings.

Predictor	**Estimate**	**Std. Error**	**z-value**	**p-value**	**Signif.**
(Intercept)	-4.4120	0.0859	-51.38	$<$2e-16	***
FCSZ1231	-0.1106	0.0120	-9.23	$<$2e-16	***
REGION202	0.2119	0.0580	3.65	0.00026	***
REGION203	0.2958	0.0513	5.76	$<$1e-08	***
REGION204	-0.0331	0.0547	-0.61	0.54512	
AGE20X	0.0883	0.0011	79.59	$<$2e-16	***
SEX2	-0.0095	0.0339	-0.28	0.78012	
RACEV2X2	0.4291	0.0494	8.68	$<$2e-16	***
RACEV2X3	0.3623	0.1784	2.03	0.04226	*
RACEV2X4	-0.4839	0.1481	-3.27	0.00109	**
RACEV2X5	-0.9313	0.1692	-5.50	$<$1e-07	***
RACEV2X6	0.1952	0.1634	1.20	0.23214	
RACEV2X10	-0.3965	0.1175	-3.38	0.00074	***
RACEV2X12	0.6143	0.1006	6.11	$<$1e-09	***

Table 7: Predictors of Chronic Disease: Logistic Regression

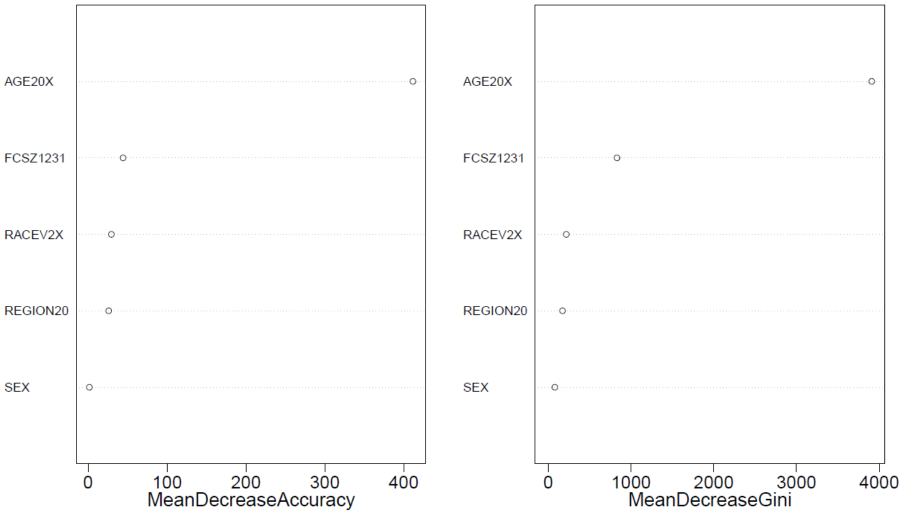


Figure 21: Random Forest Variable Importance of Demographic Features

Metric	**Logistic Regression**	**Random Forest**
Accuracy	0.8179	0.8158
Sensitivity	0.8607	0.8368
Specificity	0.7515	0.7833

Table 8: Comparison of Model Performance Metrics on Test Set

Interpretation The analysis confirms that age and family size are the strongest demographic predictors of chronic disease. Other variables such as region and race also contribute to risk differentiation. Logistic regression provides interpretable odds-based associations, while the random forest model adds predictive robustness, particularly in identifying low-risk individuals.

The finding underscore the need for targeted preventive healthcare interventions, especially for aging individuals and smaller households who may otherwise be overlooked. Public health programs and insurance providers could benefit from incorporating demographic screening tools to proactively identify and support populations at higher risk of chronic conditions. In a healthcare system where costs often escalate due to unmanaged long-term illness, such strategies can help reduce downstream expenditures and improve quality of care.

RQ 3: Distress

Summary of Findings:

- Logistic regression models revealed that insufficient sleep, low physical activity, and high smoking frequency were strongly associated with elevated psychological distress levels, as measured by the K6 scale.
- The K6SUM42 variable was binarized to define a high-distress group (top quartile) for classification purposes.
- Principal Component Analysis (PCA) was conducted to reduce dimensionality and explore latent behavioral structures. PC1 captured general wellness vs. distress, PC2 contrasted activity vs. substance use, and PC3 reflected alcohol-specific patterns.
- Four clusters were identified using K-means on the first two PCA components. These clusters exhibited distinct lifestyle profiles and were significantly associated with differences in distress levels.
- Cluster membership was found to be a statistically significant predictor of psychological distress (ANOVA p-value ¡ 0.001).
- Predictive models were developed using logistic regression, random forest, PCA-based logistic, and cluster-based logistic classifiers.

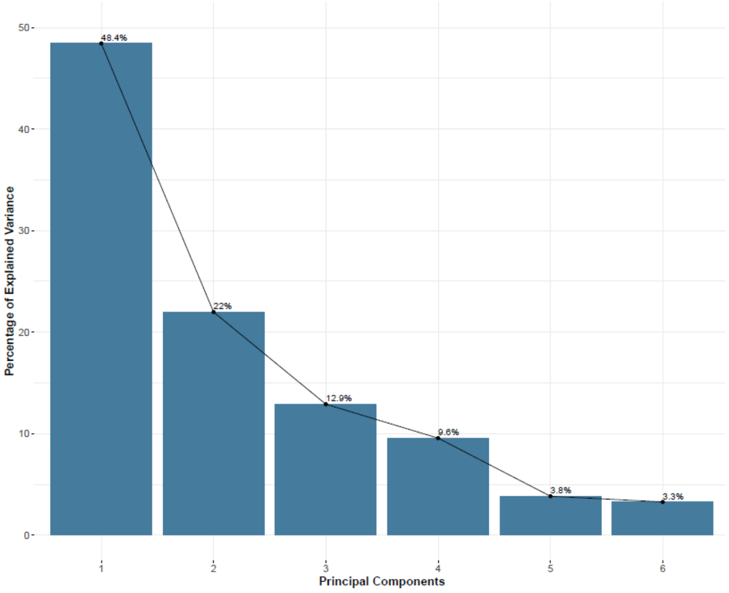


Figure 22: Scree Plot Proportion of Variance Explained by PCA

Cluster Interpretation:

Behavioral characteristics of each cluster are summarized below based on mean responses:

- **Cluster 1: Active, Low Substance Use** – Characterized by high physical activity, good sleep quality, and minimal substance use.
- **Cluster 2: Sedentary, Moderate Substance Use** – Shows lower physical activity with moderate smoking and alcohol consumption.
- **Cluster 3: Heavy Substance Users** – Distinguished by high alcohol consumption and smoking frequency.
- **Cluster 4: Poor Sleep Quality** – Marked by poor sleep quality but moderate on other measures.

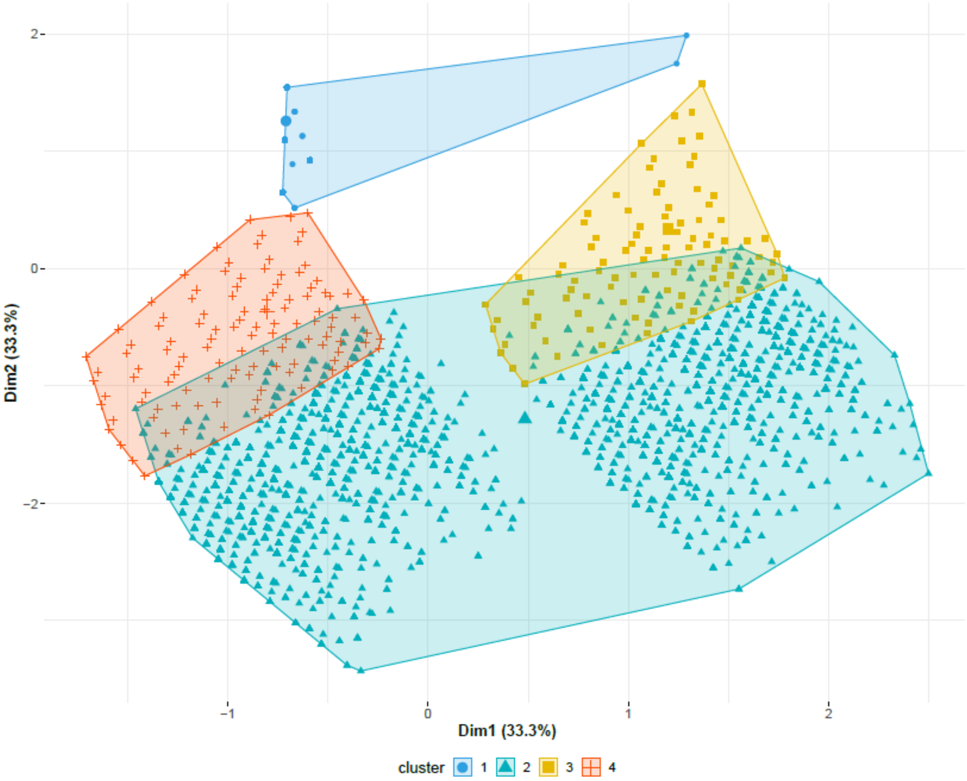


Figure 23: Clustering of Lifestyle Patterns Based on Principal Components

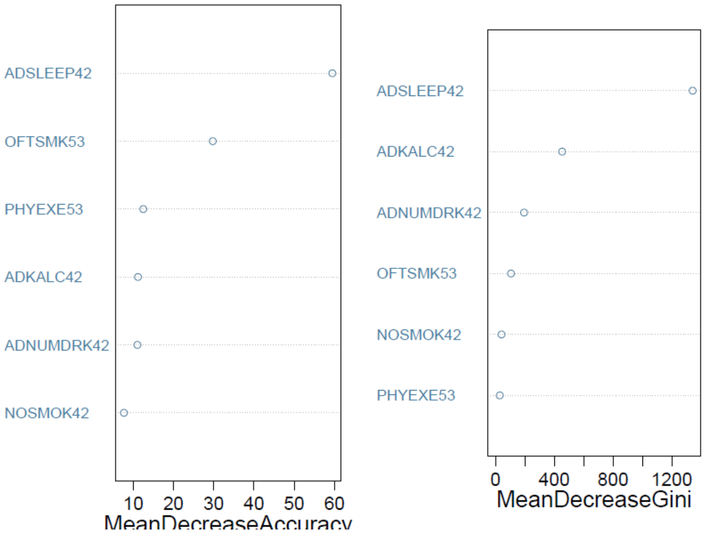


Figure 24: Variable Importance in Random Forest Based on Accuracy and Gini Impurity Reduction

Metric	Logistic Regression	Random Forest	PC-Logistic	Cluster-Logistic
AUC	0.8962	0.8511	0.8638	0.4738
Accuracy	0.8457	0.8444	0.8295	0.8086
Sensitivity	0.9368	0.9121	0.9386	1.0000
Specificity	0.4605	0.5589	0.3684	0.0000

Table 9: Comparison of Classification Performance Metrics Across Models

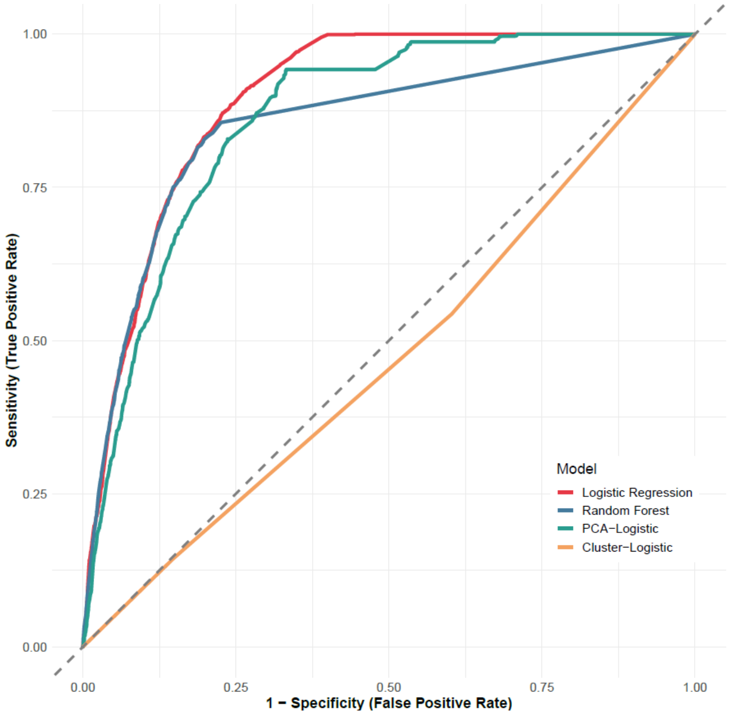


Figure 25: ROC Curve Comparison Across Models for Chronic Disease Classification

Interpretation: The results demonstrate that specific behavioral patterns are strongly associated with psychological distress. Clusters characterized by poor sleep and substance use (Clusters 3 and 4) exhibited the highest average K6 scores, while Cluster 1—those with active and health-conscious habits—showed the lowest distress. These findings support the potential for behavioral profiling to identify high-risk groups for targeted mental health interventions.

0.5 Conclusion and Limitations

This study aimed to explore the intersection of healthcare expenditures, chronic health conditions, and mental health with socioeconomic and behavioral variables using the MEPS HC-224: 2020 dataset. Across three targeted research questions, we implemented a combination of statistical and machine learning methods to extract meaningful patterns, predict outcomes, and visualize disparities.

0.5.1 Conclusion

RQ1: Expenditure Drivers in U.S. Healthcare. Expenditure Drivers in U.S. Healthcare. Analysis using multiple linear regression, ANOVA, and XGBoost identified prescription drugs, inpatient hospital care, and office-based doctor visits as the primary drivers of total annual healthcare spending. Prescription drugs had consistently high influence due to both frequency and cost, while inpatient care, though less frequent, showed the highest per-case financial impact. Office-based visits, despite their low individual cost, contributed significantly due to high utilization. These findings suggest that effective cost-control strategies should address both high-cost, low-frequency services and low-cost, high-frequency care, alongside sustained management of chronic medication use.

RQ2: Demographic Predictors of Chronic Disease. Logistic regression applied to the full dataset revealed that age and family size were the most influential demographic factors associated with the presence of chronic illness. Subsequent predictive modeling using logistic regression and random forest confirmed these variables as key predictors, with both models achieving strong classification performance. The results suggest that demographic information can serve as an effective basis for identifying high-risk groups and guiding preventive health interventions.

RQ3: Behavioral Factors and Psychological Distress. Behavioral indicators, including sleep quality, physical activity, smoking frequency, and alcohol use, were found to be strongly associated with psychological distress levels. The K6 score was transformed into a binary outcome to enable classification, and four predictive models were developed to assess model performance. In addition, principal component analysis and K-means clustering identified distinct lifestyle-based behavioral profiles, which were significantly associated with differences in psychological distress. These insights support the utility of behavioral segmentation in informing mental health policy and early intervention strategies.

0.5.2 Limitation

- The analysis relies solely on data from the 2020 MEPS Household Component file. While nationally representative, the data reflect only a single year during the COVID-19 pandemic, which may not capture typical healthcare behavior.
- All models were based on self-reported survey responses, which may introduce reporting bias, particularly for sensitive variables such as mental health indicators or substance use.
- Certain categorical variables (e.g., race, region) were included in simplified form without deeper subgroup analysis, potentially overlooking nuanced within-group variations.
- Although predictive models were implemented, model hyperparameters were tuned using basic cross-validation techniques; more rigorous optimization (e.g., grid search or Bayesian methods) could improve performance.
- The binarization of continuous variables (e.g., K6SUM42 for high distress classification) simplified analysis but potentially led to a loss of information regarding psychological severity.

Future Work Future research could extend the current analysis to a multi-year MEPS panel to evaluate healthcare trends and risk factors over time. Integrating additional external data sources, such as electronic medical records or insurance claim data, may also enhance predictive capabilities. In terms of modeling, applying advanced machine learning techniques with more thorough hyperparameter tuning may lead to better accuracy and generalizability. Lastly, subgroup analyses by race, age group, or region could uncover additional disparities and improve policy relevance.

Final Reflection This project demonstrated how nationally representative survey data can inform both statistical and predictive analyses across clinical, behavioral, and economic dimensions of healthcare. Through the integration of traditional regression techniques and modern machine learning approaches, we uncovered actionable insights into cost drivers, demographic risks, and mental health patterns. The results emphasize the potential of data-driven public health strategies to support equitable and efficient care delivery in the United States.

References

- Agency for Healthcare Research and Quality (AHRQ). *Medical Expenditure Panel Survey (MEPS) HC-224: 2020 Full Year Consolidated Data File*. Retrieved from https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-224

- Karunakaran, R., Liu, T., Zhao, M., Shah, S. S., Zhu, J., Li, H., Cai, T., & Liu, J. (2024). *Predictive interpretable analytics models for forecasting healthcare expenditures*. *Health Data Science*, Volume 4, 100053. Retrieved from https://www.sciencedirect.com/science/article/pii/S2772442524000534