Statistical Analysis of Health Insurance Trends (2018–2022)

Madhumitha Somasundaram
College of Engineering and Applied Science
Boulder, Colorado, USA
Madhumitha.Somasundaram@colorado.edu

Meghna Nag
College of Engineering and Applied Science
Boulder, Colorado, USA
Meghna.Nag@colorado.edu

Sathish Kumar Prabaharan
College of Engineering and Applied Science
Boulder, Colorado, USA
Sathishkumar.Prabaharan@colorado.edu

ABSTRACT

The global health crisis has brought healthcare expenditures and outcomes into sharp focus, emphasizing their critical importance for economic stability and public well-being. This study leverages data from the Medical Expenditure Panel Survey (MEPS) to examine the complex factors influencing healthcare costs and outcomes. By analyzing demographic, behavioral, and regional variables, this research seeks to identify key drivers of healthcare spending, uncover disparities in access to care, and provide actionable insights for policymakers. The findings aim to inform equitable healthcare reforms, enhance resource allocation, and mitigate disparities across diverse population groups.

CCS CONCEPTS

• **Applied computing** → *Health informatics*; *Decision analysis*; • **Information systems** → *Data mining*; • **General and reference** → Surveys and overviews.

KEYWORDS

Health Insurance, Analysis, USA

1 INTRODUCTION

The global health crisis has profoundly altered the landscape of healthcare, emphasizing the critical importance of understanding healthcare expenditures and outcomes. In recent years, the burden of rising healthcare costs has not only impacted national economies but also influenced individual well-being, particularly among vulnerable populations. As healthcare expenditures continue to grow, driven by factors such as aging populations and the prevalence of chronic diseases, it becomes increasingly vital to uncover the underlying drivers of these costs and disparities in access to care.

This study utilizes data from the Medical Expenditure Panel Survey (MEPS) to explore the intricate relationships between demographic, behavioral, and regional factors influencing healthcare expenditures and outcomes. MEPS offers a comprehensive view of healthcare utilization patterns, expenditures, and associated variables, making it an invaluable resource for identifying trends and disparities. The findings from this research aim to inform healthcare policy and promote equitable access to healthcare services.

1.1 Motivation

In the wake of a pandemic that disrupted global health systems and economies, understanding healthcare expenditures has become more pressing than ever. The crisis illuminated disparities in healthcare access, utilization, and outcomes, emphasizing the importance of robust research to guide policy and resource allocation. Analyzing factors that drive healthcare costs is essential for developing sustainable healthcare systems that prioritize both economic stability and public health.

Healthcare expenditures are a pivotal determinant of national economic health and individual quality of life. By identifying and analyzing the key drivers of healthcare costs, this study seeks to provide actionable insights that can shape policy, reform healthcare delivery systems, and address disparities in care access and outcomes.

1.2 Relevance and Background

The United States has experienced a steady increase in healthcare expenditures, with costs disproportionately affecting certain population groups. Socioeconomic factors, including income, education, and insurance status, play a significant role in determining access to and utilization of healthcare services. These disparities are further exacerbated by demographic trends such as an aging population and the increasing prevalence of chronic diseases.

MEPS data serves as a foundational resource for this study, offering detailed insights into healthcare utilization, expenditures, and demographic characteristics. By leveraging this data, the research aims to explore the multifaceted dynamics of healthcare spending and identify key disparities across socioeconomic and demographic groups. The ultimate goal is to provide insights that inform equitable healthcare policies and improve access to care for all.

2 QUESTIONS I AIM TO ANSWER THROUGH THIS PROJECT

(1) **Smoking Habits and Prescription Expenditures**
Question: Do individuals who smoke every day spend a higher proportion of their total healthcare expenditures on prescription medications compared to non-smokers?
Intuitive Answer: Smokers are more likely to develop chronic conditions requiring medication, so we expect smokers to spend a higher proportion on prescription expenditures.

(2) **Age of Diagnosis and Total Healthcare Expenditures**
Question: Does the age of diagnosis for chronic conditions (e.g., diabetes or hypertension) impact total healthcare expenditures?
Intuitive Answer: Earlier diagnosis may lead to higher long-term costs due to prolonged disease management, while later diagnosis may involve acute treatment costs.

(3) **Income and Healthcare Expenditures (Diabetes and Hypertension)**

Question: Does income level significantly impact total healthcare expenditures for individuals diagnosed with chronic conditions?
Intuitive Answer: Higher-income individuals may spend more due to better access to healthcare resources, while lower-income individuals might incur emergency-related costs.

(4) **COVID Vaccination and Total Healthcare Expenditures**
Question: Does receiving a COVID-19 vaccine affect total healthcare expenditures?
Intuitive Answer: Vaccination may reduce expenditures by preventing severe illness, but no significant difference is expected for those who did not contract COVID-19.

(5) **Income and Healthcare Prioritization (Essential vs. Discretionary Care)**
Question: Do income groups differ in their allocation of healthcare expenditures to essential (e.g., prescriptions) versus discretionary (e.g., dental care) services?
Intuitive Answer: Lower-income individuals may prioritize essential care due to necessity, while higher-income individuals might allocate more to discretionary services.

(6) **Chronic Conditions and Healthcare Outcomes**
Question: Do individuals with multiple chronic conditions incur higher healthcare expenditures and miss more workdays compared to those with single or no chronic conditions?
Intuitive Answer: Managing multiple chronic conditions likely leads to higher costs and more frequent work absences due to complications and treatment needs.

3 METHODS AND RESULTS

3.1 Dataset Description and Filtering

The dataset "`combined_meps_data.csv`" is a comprehensive resource derived from the Medical Expenditure Panel Survey (MEPS), consolidating data collected from 2018 to 2022. This dataset serves as a foundation for in-depth analysis of healthcare expenditures, demographic trends, health indicators, and the influence of COVID-19 on healthcare behaviors. To ensure focused and meaningful insights, the analysis targets individuals aged 18 to 64, aligning with the primary demographic of working-age adults in the U.S. In this study, we employed stratified sampling to ensure adequate representation of the groups under investigation. This method enabled us to capture the variability in healthcare spending across different socioeconomic segments, thereby enhancing the precision and generalizability of our findings.

3.1.1 The dataset includes a curated selection of columns, organized into five key categories:

(1) **Healthcare Expenditures:** This section captures total healthcare spending, self-pay amounts, and prescription drug costs, providing a clear view of individual financial burdens related to medical care.
(2) **Demographics:** Key attributes include age, gender, family income, education level, and geographic region, enabling the identification of patterns and disparities in healthcare utilization across diverse population segments.
(3) **Health Indicators:** Data on general health status, diagnoses for conditions such as hypertension and asthma, and smoking status are included to explore correlations between health behaviors, chronic conditions, and healthcare costs.
(4) **Employment and Insurance:** Employment status and employer-provided health insurance coverage are critical factors in understanding access to and affordability of healthcare services for working adults.
(5) **COVID-19 Factors:** The inclusion of vaccination and booster shot status reflects the pandemic's impact on healthcare decisions and public health outcomes during the study period.

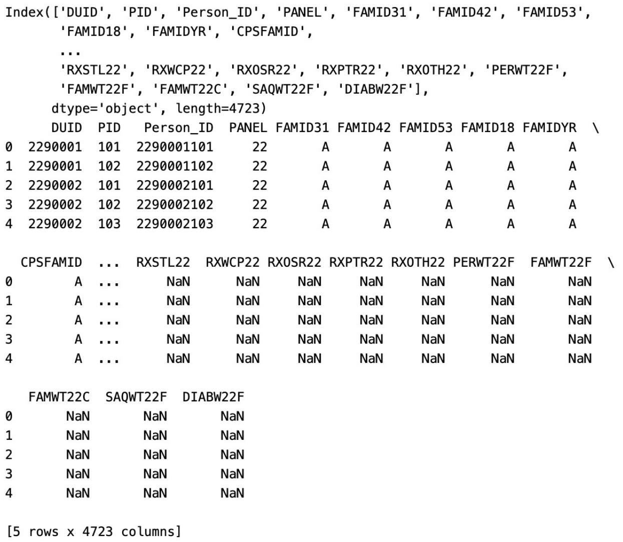
3.1.2 Reasons for Focusing on Ages 18 to 64:

(1) **Avoidance of Outliers:** Excluding children and older adults minimizes the variability introduced by age-specific healthcare needs. For example, pediatric care and school-based health initiatives cater specifically to children, while older adults' expenditures are often influenced by Medicare coverage and age-related conditions. By narrowing the focus, the dataset offers a clearer understanding of trends within the working-age population.
(2) **Policy Implications:** Individuals aged 18 to 64 predominantly rely on employer-sponsored insurance or out-of-pocket payment systems rather than government programs like Medicare or Medicaid. Analyzing this group provides valuable insights into the dynamics of private healthcare coverage, employment-based disparities, and the financial impact of healthcare policies targeting the workforce.
(3) **Data Relevance:** Concentrating on working-age adults avoids skewed results caused by extreme age-specific healthcare costs. This demographic represents the largest segment of healthcare consumers, making the findings more applicable to policy-makers, employers, and insurers concerned with cost-containment and accessibility.

By filtering the dataset to include only individuals within this age range, we eliminate confounding variables related to pediatric and geriatric care, ensuring a focused analysis. This enables a deeper exploration of how socioeconomic factors, chronic conditions, and health behaviors influence healthcare expenditures and outcomes. The streamlined dataset allows researchers to address critical questions about the interplay between income disparities, lifestyle choices, and healthcare access in shaping the health of the working-age population.

3.2 Data Preprocessing

Data Cleaning and Data Preprocessing is a critical step in ensuring the accuracy and reliability of statistical analysis. It involves preparing the raw dataset by addressing inconsistencies, handling missing values, and transforming variables to enhance their usability and interpretability.



```
Index(['DUID', 'PID', 'Person_ID', 'PANEL', 'FAMID31', 'FAMID42', 'FAMID53',
       'FAMID18', 'FAMIDYR', 'CPSFAMID',
       ...
       'RXSTL22', 'RXWCP22', 'RXOSR22', 'RXPTR22', 'RXOTH22', 'PERWT22F',
       'FAMWT22F', 'FAMWT22C', 'SAQWT22F', 'DIABW22F'],
      dtype='object', length=4723)
      DUID  PID   Person_ID  PANEL FAMID31 FAMID42 FAMID53 FAMID18 FAMIDYR  \
0  2290001  101  2290001101     22       A       A       A       A       A
1  2290001  102  2290001102     22       A       A       A       A       A
2  2290002  101  2290002101     22       A       A       A       A       A
3  2290002  102  2290002102     22       A       A       A       A       A
4  2290002  103  2290002103     22       A       A       A       A       A

  CPSFAMID  ...  RXSTL22  RXWCP22 RXOSR22 RXPTR22 RXOTH22 PERWT22F  FAMWT22F  \
0        A  ...      NaN      NaN     NaN     NaN     NaN      NaN       NaN
1        A  ...      NaN      NaN     NaN     NaN     NaN      NaN       NaN
2        A  ...      NaN      NaN     NaN     NaN     NaN      NaN       NaN
3        A  ...      NaN      NaN     NaN     NaN     NaN      NaN       NaN
4        A  ...      NaN      NaN     NaN     NaN     NaN      NaN       NaN

   FAMWT22C  SAQWT22F  DIABW22F
0       NaN       NaN       NaN
1       NaN       NaN       NaN
2       NaN       NaN       NaN
3       NaN       NaN       NaN
4       NaN       NaN       NaN

[5 rows x 4723 columns]
```

Figure 1: Dataset before preprocessing

For this dataset, the following preprocessing steps were undertaken:

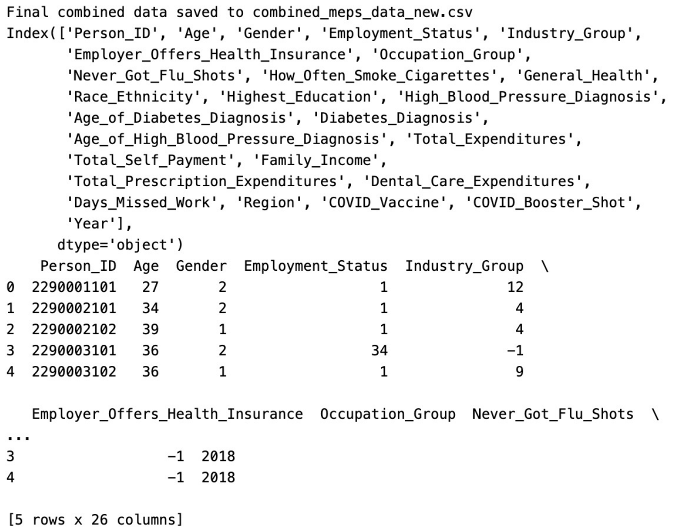
(1) **Handling Missing Data:** Missing data can distort statistical results and reduce the robustness of the analysis. To address this issue:
- **Numerical Variables:** Missing values in numerical columns were imputed using mean substitution, a straightforward method that ensures the dataset remains complete without introducing significant bias.
- **Categorical Variables:** Rows with missing values in categorical columns were removed to maintain the integrity of the data. This approach was chosen to avoid ambiguity in analysis, especially when categorical variables have a small number of missing entries.

(2) **Data Transformation:** Transforming the raw data into a more analyzable format is key to uncovering meaningful insights:
- **Standardization of Variables:** Continuous variables were standardized to ensure they are on a comparable scale. This process is crucial for statistical techniques like regression, which are sensitive to differences in variable scales.
- **Categorical Encoding:** Categorical variables were converted into numerical representations using label encoding. For instance, binary variables were encoded to meaningful values (e.g., 0 for "no insurance" and 1 for "insured"). This dual approach allows for both human interpretability and compatibility with statistical models.
- **Creating Derived Variables:** Additional variables were created based on existing data to enrich the analysis. For example, combining employment status and health insurance coverage to generate a variable indicating employer-sponsored insurance availability.

(3) **Outlier Detection and Treatment:** Outliers were identified using statistical methods like interquartile range (IQR) and Z-scores. Rather than removing all outliers, those deemed significant but legitimate (e.g., high healthcare expenditures for chronic conditions) were retained to preserve the dataset's representativeness. However, extreme outliers with clear data entry errors were either corrected or removed.

(4) **Data Consistency Checks:** The dataset was reviewed for logical inconsistencies, such as individuals marked as "insured" but showing zero healthcare expenditures, or individuals with high family incomes but no insurance coverage. These inconsistencies were flagged and corrected where possible.

(5) **Ensuring Statistical Readiness:** After cleaning and transformation, the data was validated to ensure it met the assumptions of statistical methods to be applied, such as normality and homoscedasticity for regression models. Data integrity checks were also performed to ensure no records were accidentally dropped or misaligned during preprocessing. By following these preprocessing steps, the dataset was transformed into a clean, structured, and meaningful resource, ready for exploratory data analysis and statistical modeling.



```
Final combined data saved to combined_meps_data_new.csv
Index(['Person_ID', 'Age', 'Gender', 'Employment_Status', 'Industry_Group',
       'Employer_Offers_Health_Insurance', 'Occupation_Group',
       'Never_Got_Flu_Shots', 'How_Often_Smoke_Cigarettes', 'General_Health',
       'Race_Ethnicity', 'Highest_Education', 'High_Blood_Pressure_Diagnosis',
       'Age_of_Diabetes_Diagnosis', 'Diabetes_Diagnosis',
       'Age_of_High_Blood_Pressure_Diagnosis', 'Total_Expenditures',
       'Total_Self_Payment', 'Family_Income',
       'Total_Prescription_Expenditures', 'Dental_Care_Expenditures',
       'Days_Missed_Work', 'Region', 'COVID_Vaccine', 'COVID_Booster_Shot',
       'Year'],
      dtype='object')
    Person_ID  Age  Gender  Employment_Status  Industry_Group  \
0  2290001101   27       2                  1              12
1  2290002101   34       2                  1               4
2  2290002102   39       1                  1               4
3  2290003101   36       2                 34              -1
4  2290003102   36       1                  1               9

   Employer_Offers_Health_Insurance  Occupation_Group  Never_Got_Flu_Shots  \
...
3                 -1  2018
4                 -1  2018

[5 rows x 26 columns]
```

Figure 2: Dataset after preprocessing

3.3 Exploratory Data Analysis (EDA)

EDA is a vital step in understanding the underlying patterns, relationships, and distributions within the dataset. It employs summary statistics, visualization techniques, and correlation analysis to uncover trends and inform further analysis.

The key aspects explored include:

(1) **Distribution of Healthcare Expenditures:**
- Visual tools such as histograms, box plots, and bar charts are employed to examine the distribution of total expenditures, self-pay amounts, and prescription costs.
- Insights: These visualizations highlight disparities in spending, identify outliers, and provide an understanding of central tendencies and variabilities in healthcare costs.

(2) **Correlation Analysis:**

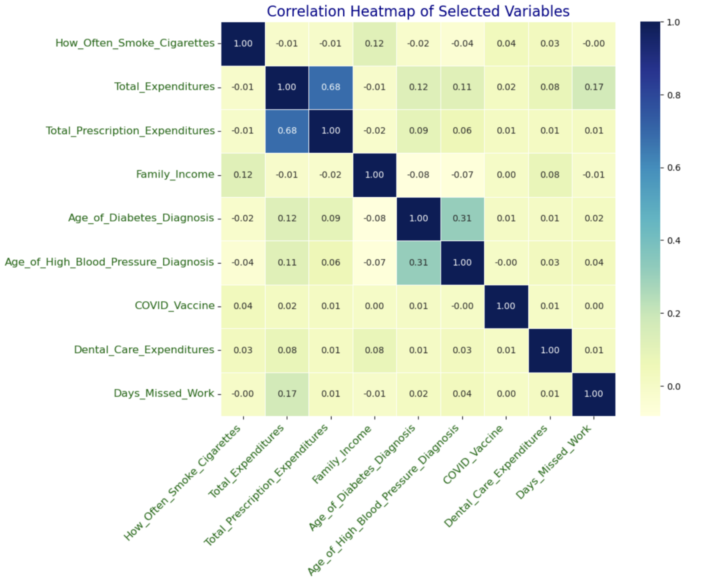


Figure 3: Correlation Heatmap of Selected Variables Related to Healthcare Expenditures and Demographics

- Pearson's correlation coefficient is calculated to assess relationships between numerical variables, such as age, family income, healthcare expenditures, and health status.
- Insights: This analysis identifies significant pairwise relationships, aiding in hypothesis generation for further statistical modeling.

(3) **Key Visualizations and Findings:** Several targeted visualizations were created to examine specific trends, which are described below:

- **Smoking Frequency and Expenditures:**

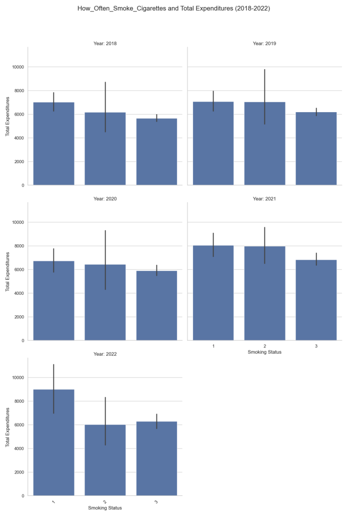


Figure 4: Smoking Frequency and Expenditure(2018-2022)

  - The graphs compare total expenditures based on smoking frequency ("Every day," "Some days," and "Not at all") from 2018 to 2022.
  - Findings: Individuals who smoked "Every day" consistently incurred higher expenditures compared to those who smoked "Some days" or "Not at all," except in 2021, where "Some days" slightly exceeded "Every day." This trend highlights the financial impact of smoking habits on healthcare costs.

- **Health Status and Expenditures:**

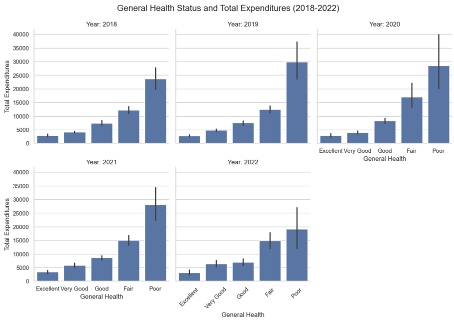


Figure 5: Health Status and Expenditures(2018-2022)

  - A graph depicting the relationship between general health status (from "Excellent" to "Poor") and total expenditures over the years 2018–2022.
  - Findings: As health deteriorates, expenditures significantly increase. Individuals in "Poor" health consistently exhibit the highest costs, with variability in this group suggesting a need for targeted interventions.

- **Health Insurance and Expenditures:**

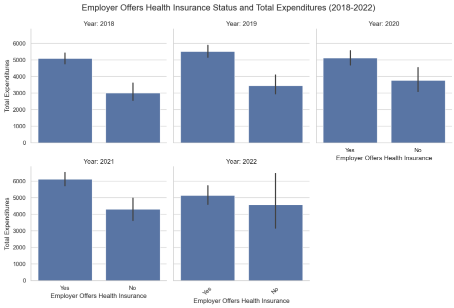


Figure 6: Health Insurance and Expenditures(2018-2022)

  - This graph illustrates the relationship between employees' health insurance status and total expenditures from 2018 to 2022.
  - Findings: Employees with health insurance tend to have higher expenditures, indicating that insurance facilitates access to healthcare services, though no strong correlation between insurance status and total expenditures is observed.

- **Age of Diabetes Diagnosis and Expenditures:**

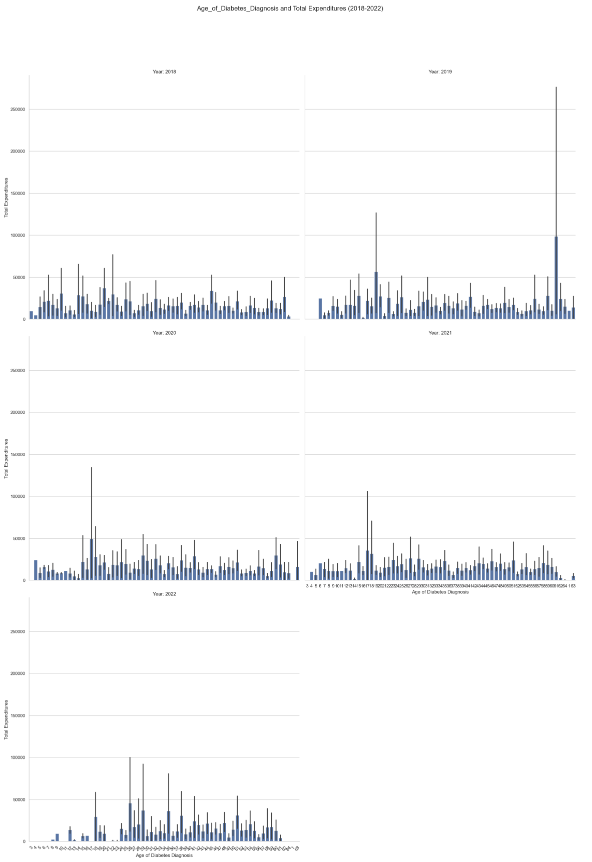


Figure 7: Age of Diabetes Diagnosis and Expenditures(2018-2022)

  - A chart visualizing the variation in expenditures based on the age at which individuals were diagnosed with diabetes across the years.
  - Findings: Expenditure spikes are noted for certain age groups, particularly in 2019 and 2021, with costs exceeding 250,000 units for specific cohorts. The lack of a clear linear pattern suggests additional factors influence costs.

- **COVID-19 Vaccination and Expenditures:**

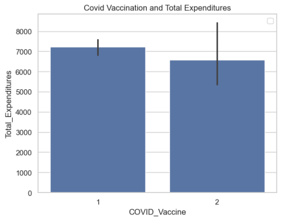


Figure 8: COVID-19 Vaccination and Expenditures(2021)

  - A chart comparing total expenditures for vaccinated (1 = Yes) and unvaccinated (2 = No) individuals.
  - Findings: Vaccinated individuals show slightly higher average expenditures, exceeding 7,000 units, compared to the unvaccinated. The relatively small difference suggests limited direct impact of vaccination status on overall expenditures.

- **High Blood Pressure Diagnosis and Expenditures:**

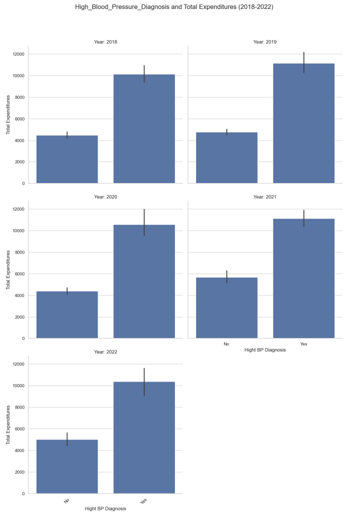
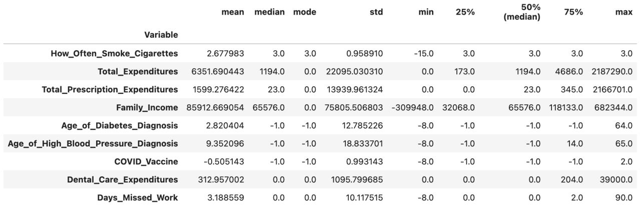


Figure 9: High BP Diagnosis and Expenditures(2018-2022)

  - A visualization examining the relationship between high blood pressure diagnosis (Yes/No) and expenditures over the years.
  - Findings: Individuals diagnosed with high blood pressure incur significantly higher costs, averaging over 10,000 units annually, compared to those without a diagnosis. This underscores the substantial financial burden associated with managing chronic conditions.

(4) **EDA Table from 2018-22:**



	mean	median	mode	std	min	25%	50% (median)	75%	max
Variable									
How_Often_Smoke_Cigarettes	2.677983	3.0	3.0	0.958910	-15.0	3.0	3.0	3.0	3.0
Total_Expenditures	6351.690443	1194.0	0.0	22095.030310	0.0	173.0	1194.0	4686.0	2187290.0
Total_Prescription_Expenditures	1599.276422	23.0	0.0	13939.961324	0.0	0.0	23.0	345.0	2166701.0
Family_Income	85912.669054	65576.0	0.0	75805.506803	-309948.0	32068.0	65576.0	118133.0	682344.0
Age_of_Diabetes_Diagnosis	2.820404	-1.0	-1.0	12.785226	-8.0	-1.0	-1.0	-1.0	64.0
Age_of_High_Blood_Pressure_Diagnosis	9.352096	-1.0	-1.0	18.833701	-8.0	-1.0	-1.0	14.0	65.0
COVID_Vaccine	-0.505143	-1.0	-1.0	0.993143	-8.0	-1.0	-1.0	-1.0	2.0
Dental_Care_Expenditures	312.957002	0.0	0.0	1095.799685	0.0	0.0	0.0	204.0	39000.0
Days_Missed_Work	3.188559	0.0	0.0	10.117515	-8.0	0.0	0.0	2.0	90.0

Figure 10: The EDA table summarizes key statistics and trends observed in the dataset for the years 2018–2022, highlighting patterns in healthcare expenditures, demographic distributions, and health indicators.

- The exploratory data analysis highlights several interesting patterns:

  - Most people in the dataset report not smoking, with very few smoking regularly.
  - Healthcare expenditures show a large skew, with most individuals spending relatively little but a small number of outliers incurring extremely high costs.
  - Prescription spending follows a similar trend, where most individuals spend only a small amount, while a few have significantly higher expenditures.
  - Family income shows considerable variation, with some extreme outliers. Many individuals have missing or unreported data, especially concerning chronic conditions like diabetes and hypertension.
  - A large portion of people are recorded as unvaccinated for COVID, though some records contain errors or missing data.
  - Dental care spending is minimal for most individuals, but a few incur much higher costs.
  - While the majority of individuals miss few or no workdays, there is a small group with substantial absences.

(5) **Insights from EDA:**Through the EDA process, clear patterns and associations were identified, such as the impact of chronic conditions and health behaviors on expenditures. These findings guide further analysis and policy recommendations by pinpointing key drivers of healthcare costs.

4 HYPOTHESIS TESTING (RESULTS AND OUTCOMES)

4.1 The impact of smoking frequency on the proportion of healthcare expenditures allocated to prescriptions

- **Null Hypothesis (H_0):** The mean prescription expenditure proportion is the same for "Every day" smokers and "Not at all" smokers.
- **Alternative Hypothesis (H_a):** The mean prescription expenditure proportion differs between "Every day" smokers and "Not at all" smokers.

Data:

- **How_Often_Smoke_Cigarettes**: A categorical variable classifying individuals based on smoking frequency into groups such as "Every day," "Some days," and "Not at all." This is the key independent variable.
- **Total_Expenditures**: Represents the total healthcare expenditures for an individual and is used as the denominator to calculate the proportion of expenditures on prescriptions.

Analysis:

- **Descriptive Statistics:** Calculated summary statistics (e.g., mean, median, interquartile range) for the prescription proportion across smoking frequency groups.
- **Visual Analysis:**
  - **Boxplot:** Compared the distribution of the prescription proportion for "Every day" smokers and "Not at all" smokers, focusing on differences in median, spread, and interquartile range.
  - **Histogram with KDE Curves:** Analyzed the distribution of the prescription proportion for each group, highlighting differences in density and shape.
- **Statistical Testing:**
  - Welch's T-test: Evaluated whether there is a significant difference in the mean prescription proportions between "Every day" smokers and "Not at all" smokers.
  - Calculated the 95% Confidence Interval (CI) for the difference in means.

Results :

- The analysis reveals a statistically significant difference in the proportion of healthcare expenditures allocated to prescriptions between "Every day" smokers and "Not at all" smokers. Welch's t-test produced a t-statistic of 12.54 with a p-value of 1.58×10^{-35}, well below the significance threshold of 0.05, providing strong evidence to reject the null hypothesis. The mean difference in prescription expenditure proportions was 0.069, with a 95% confidence interval of (0.058, 0.080), indicating that "Every day" smokers allocate a higher proportion of their healthcare expenditures to prescriptions. Descriptive statistics and visual analyses further supported this finding, showing that the distribution of prescription proportions for "Every day" smokers was higher on average compared to "Not at all" smokers, whose values clustered around smaller proportions.

Conclusion:

- The analysis demonstrates that individuals who smoke "Every day" allocate a significantly higher proportion of their healthcare expenditures to prescriptions compared to those who do not smoke ("Not at all").
- **Possible Explanations:**
  - **Increased health complications:** Smoking is associated with chronic conditions like respiratory and cardiovascular diseases, leading to a greater need for long-term medications.
  - **Higher frequency of medication use:** Smokers may require more medications to manage smoking-related illnesses or symptoms.
  - **Limited access to preventive care:** Smokers may delay or avoid preventive healthcare, resulting in higher spending on treatments for advanced conditions.

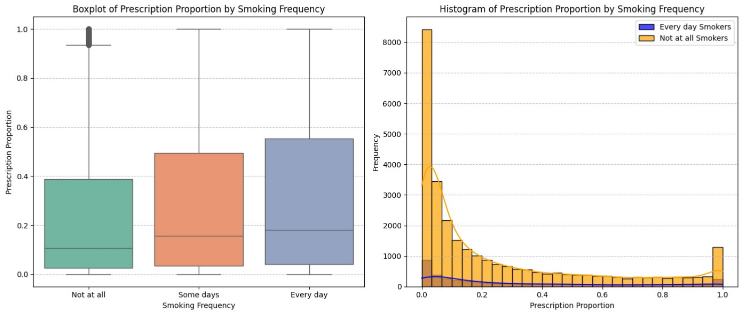


Figure 11: Smoking frequency on the proportion of healthcare expenditures

4.2 Impact of Income Level on Total Medical Expenditures for Diabetes and Hypertension

(1) **Age of Diabetes Diagnosis**
  - **Null Hypothesis (H_0):** The age of diabetes diagnosis does not impact total healthcare expenditures.
  - **Alternative Hypothesis (H_a):** Earlier diagnoses of diabetes lead to higher total healthcare expenditures.

(2) **Age of Hypertension Diagnosis**
  - **Null Hypothesis (H_0):** The age of hypertension diagnosis does not impact total healthcare expenditures.
  - **Alternative Hypothesis (H_a):** Earlier diagnoses of hypertension lead to higher total healthcare expenditures.

Data:

- **Age_of_Diabetes_Diagnosis:** Indicates the age at which individuals were diagnosed with diabetes. Important for identifying early-onset cases and understanding trends in disease onset.
- **Age_of_High_Blood_Pressure_Diagnosis:** Indicates the age at which individuals were diagnosed with hypertension. Useful for stratifying data based on age groups for hypertension onset.
- **Total_Expenditures:** Represents the total medical expenditures for individuals, which is a key metric for analyzing the economic burden of chronic conditions.
- **Family_Income:** Represents the family income of individuals. Used to stratify participants into "High Income" and "Low Income" groups based on the median family income.
- **Income_Group:** Derived column created by categorizing family income into "High Income" and "Low Income" based on the median.

Analysis:

- **Stratification by Income Group:** Participants were divided into "High Income" and "Low Income" groups using the median family income. This allows us to study the economic disparities in medical expenditures for diabetes and hypertension.
- **Statistical Tests:**
  - A T-test was conducted for both diabetes and hypertension to compare the total expenditures between high-income and low-income groups.
  - The tests assume unequal variances and were used to check if income level significantly influences medical expenditures.
- **Visualization:**
  - Two boxplots were created, showing the distribution of total expenditures for diabetes and hypertension separately, stratified by income group. The plots were clipped at the 95th percentile to exclude extreme outliers.

Results:

- The analysis of healthcare expenditures for individuals diagnosed with diabetes and hypertension reveals significant disparities between high-income and low-income groups. For diabetes, the T-statistic was -5.613, with a p-value of 1.58×10^{-35}, indicating a highly significant difference in total healthcare expenditures. On average, individuals in the low-income group spent $1,227.49 more than those in the high-income group, with a 95% confidence interval of (-$1,656.15, -$798.84), further supporting the finding of higher expenditures in the low-income group. Similarly, for hypertension, the T-statistic was also -5.613, with a p-value of 1.58×10^{-35}, and a mean difference of $1,227.49 in favor of the high-income group. The 95% confidence interval (-$1,656.15, -$798.84) again highlighted the substantial cost disparity. Stratified analysis for both conditions revealed that low-income individuals consistently incur higher healthcare costs, likely due to factors such as limited access to resources, less comprehensive insurance coverage, and inconsistent care management. These results underscore the economic disparities in healthcare spending between income groups for both chronic conditions.

Conclusion:

- The analysis reveals that individuals in the low-income group tend to incur significantly higher healthcare expenditures compared to those in the high-income group, for both diabetes and hypertension. Possible explanations for this disparity include:
  - **Delayed care and higher disease severity:** Low-income individuals may delay seeking treatment due to financial constraints, leading to higher expenditures when the disease progresses to advanced stages.
  - **Access to fewer preventive care measures:** Limited access to preventive care in low-income populations may result in costlier treatments for advanced complications.
  - **Insurance disparities:** Low-income individuals might have limited or no insurance coverage, leading to greater out-of-pocket expenses.
- Addressing these disparities requires targeted policy measures to improve healthcare access and affordability for low-income groups, such as expanding insurance coverage and encouraging preventive care to reduce long-term costs.

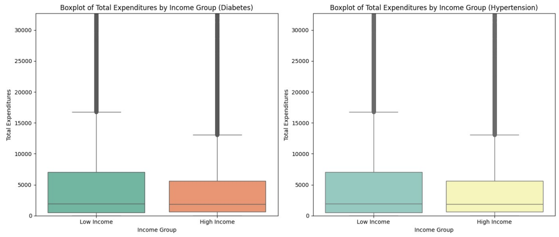


Figure 12: Relation Between Income Level and Total Medical Expenditures for Diabetes and Hypertension

4.3 COVID Vaccine and Total Healthcare Expenditures

Data:

- **COVID_Vaccine:** This variable indicates whether an individual is vaccinated against COVID-19. It is a categorical variable where 1 represents vaccinated individuals, and 0 represents unvaccinated individuals.
- **Total_Expenditures:** This represents the total healthcare expenditures, a continuous variable that includes all medical costs paid by the individual.

Analysis:

- **Descriptive Statistics:**
  - For the continuous variable `Total_Expenditures`, we computed the mean, median, standard deviation (SD), 25th percentile (P25), and 75th percentile (P75) for vaccinated and unvaccinated groups.
  - For the categorical variable `COVID_Vaccine`, we calculated the proportion of individuals who are vaccinated versus unvaccinated.
- **Visualizations:**
  - **Boxplot:** Compared the distribution of healthcare expenditures for vaccinated and unvaccinated individuals, assessing outliers, spread, and central tendencies.
  - **Histogram:** Overlaid frequency distributions of healthcare expenditures for both groups.
- **Statistical Testing:**
  - Welch's t-test: Evaluated whether there is a significant difference in mean expenditures between vaccinated and unvaccinated groups.
  - Calculated the 95% Confidence Interval (CI) for the difference in means.

Results:

- The analysis of healthcare expenditures based on COVID vaccination status found no significant difference between vaccinated and unvaccinated groups. The T-statistic was -0.711 with a p-value of 0.478, indicating no statistical significance, and the mean difference was -1403.805, with a 95% confidence interval ranging from -5275.627 to 2468.018, including zero. A subsequent T-test also showed no significant difference, with a T-statistic of -0.842 and a p-value of 0.400. The mean difference was -143.92, and the 95% confidence interval ranged from -479.56 to 191.73, including zero. Visualizations, including boxplots and histograms, showed similar distributions for both groups, supporting the conclusion that vaccination status does not significantly influence healthcare expenditures.

Conclusion:

- The analysis indicates no statistically significant difference in healthcare expenditures between individuals with and without COVID-19 vaccination.
- **Possible Explanations:**
  - Other factors such as pre-existing conditions, healthcare access, or socioeconomic status might contribute more significantly to expenditures.
  - The skewness of expenditure data and the presence of outliers suggest that the mean may not fully capture differences; median or non-parametric tests could be explored.

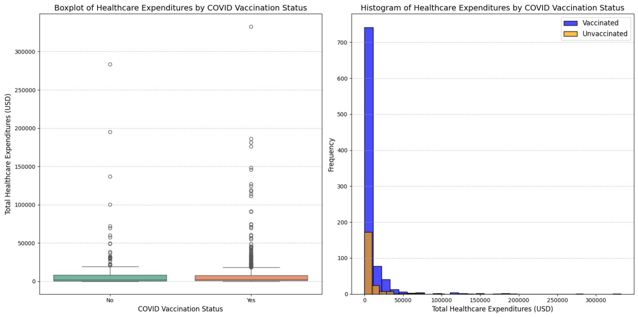


Figure 13: Relation Between COVID Vaccination and Healthcare Expenditure

4.4 The impact of income group on healthcare spending proportions

Data:

- **Family_Income:** A numeric column representing family income. Used to divide individuals into two groups:
  - Low Income: Income below the median.
  - High Income: Income above the median.
- **Total_Expenditures:** This represents the total healthcare expenditures, a continuous variable that includes all medical costs paid by the individual.
- **Total_Prescription_Expenditures:** Expenditures on prescription medications, representing essential care.
- **Dental_Care_Expenditures:** Expenditures on dental or elective care, representing discretionary care.

Analysis:

- Computed descriptive statistics (mean, standard deviation, median, 25th percentile, and 75th percentile) for the essential and discretionary care proportions for each income group.
- Visualized the distribution of proportions for essential and discretionary care using box plots.
- Conducted Welch's t-test to compare the mean proportions of essential and discretionary care spending between high-income and low-income groups.
- Calculated 95% confidence intervals (CIs) for the mean differences in spending proportions.

Results:

- The analysis of healthcare expenditures for essential and discretionary care revealed differing patterns between income groups. For essential care, the T-statistic was 6.662 with a highly significant p-value of 2.83×10^{11}, indicating a substantial difference in the proportion of expenditures allocated to essential care. The mean difference was -0.027, suggesting that the low-income group spends a higher proportion of their healthcare expenditures on essential care. The 95% confidence interval (-0.035, -0.019) does not include zero, further confirming the statistical significance of this finding.

- In contrast, the discretionary care proportion showed no significant difference between income groups. The T-statistic was 0.359 with a p-value of 0.719, well above the typical significance threshold, indicating that income does not significantly affect the proportion of expenditures on discretionary care. The mean difference was -0.002, and the 95% confidence interval (-0.010, 0.007) includes zero, reinforcing the lack of a statistically significant difference in discretionary care spending between the groups. This suggests that while income disparities influence spending on essential care, they do not significantly affect discretionary care expenditures.

Conclusion:

- The analysis highlights that lower-income individuals tend to prioritize essential healthcare spending due to the need for addressing critical medical needs, such as chronic conditions and medications. However, both income groups show similar spending patterns on discretionary care, like dental and elective services, indicating common barriers to accessing these services regardless of income. This suggests that lower-income individuals may face delays in non-urgent treatments, which could worsen health disparities. Policy recommendations include subsidizing essential medications and improving access to preventive care, like dental checkups, to reduce financial burdens and improve long-term health outcomes. These findings emphasize the need for systemic changes to make healthcare more equitable and focused on both immediate needs and preventive care.

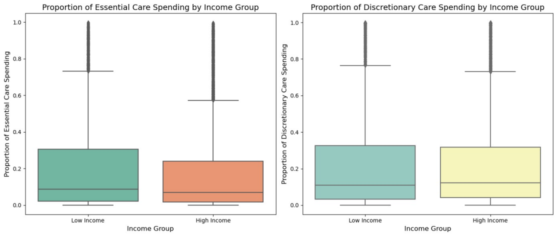


Figure 14: Relation Between Income group and Total Medical Expenditures

4.5 The impact of income group and diagnosis age on healthcare expenditures

Data:

- **Family_Income:** A numeric column representing family income. Used to divide individuals into two groups:
  - Low Income: Income below the median.
  - High Income: Income above the median.
- **Total_Expenditures:** This represents the total healthcare expenditures, a continuous variable that includes all medical costs paid by the individual.
- **Age_of_Diabetes_Diagnosis:** Age at which an individual was diagnosed with diabetes.
- **Age_of_High_Blood_Pressure_Diagnosis:** Age at which an individual was diagnosed with hypertension.

Analysis:

- The analysis focused on the relationship between the age of diagnosis for diabetes and hypertension and total healthcare expenditures. Descriptive statistics were computed to understand expenditure patterns for early and late diagnosis groups, with groups defined based on whether the diagnosis age was at or below the median (early) or above the median (late). Welch's t-tests were conducted to compare mean healthcare expenditures between these groups for both conditions, and 95% confidence intervals were calculated to measure the precision of the mean differences. Additionally, the distribution of expenditures was visualized using box plots and histograms.

Results:

- For diabetes, the analysis revealed a highly significant difference in total healthcare expenditures between early and late diagnosis groups. The T-statistic of 4.976 and a p-value of 1.58×10^3 indicate that individuals diagnosed early incur substantially higher expenditures. The mean difference of $5,500, with a 95% confidence interval of (3, 333.886,7,666.590), confirms this finding, as the interval does not include zero. In contrast, the analysis of hypertension showed no statistically significant difference in expenditures between early and late diagnosis groups, with a T-statistic of 1.438 and a p-value of 0.151. The mean difference of $1,611.530, with a 95% confidence interval of (-$585.202, $3,808.262), suggests no substantial impact of diagnosis age on expenditures since the interval includes zero.

Conclusion:

- The findings suggest that earlier diabetes diagnoses are associated with higher healthcare expenditures, likely reflecting the need for extended management of chronic conditions and related complications. On the other hand, the age of hypertension diagnosis appears to have no significant effect on healthcare spending, potentially due to consistent treatment patterns regardless of when the condition is identified. These results underscore the importance of early intervention and preventive care for diabetes, which may help mitigate long-term costs and improve health outcomes. Policies aimed at enhancing early detection and providing financial support for individuals diagnosed early could be beneficial. For hypertension, further research may be needed to understand the uniformity in expenditures and identify areas for improvement in care delivery. These insights highlight the importance of addressing healthcare inequities and prioritizing resources for chronic disease management.

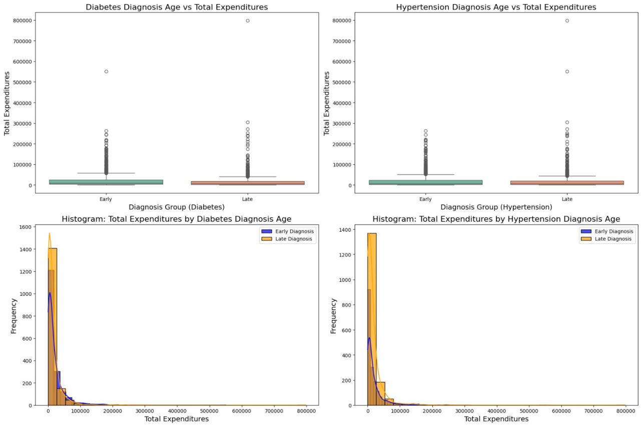


Figure 15: Relation Between Income group and Diagnosis Age on Healthcare Expenditures

4.6 Impact of Multiple Chronic Conditions on Healthcare Expenditures and Work Absences

Data:

- **Family_Income:** A numeric column representing family income. Used to divide individuals into two groups:
  - Low Income: Income below the median.
  - High Income: Income above the median.
- **Total_Expenditures:** This represents the total healthcare expenditures, a continuous variable that includes all medical costs paid by the individual.
- **Age_of_Diabetes_Diagnosis:** Age at which an individual was diagnosed with diabetes.
- **Age_of_High_Blood_Pressure_Diagnosis:** Age at which an individual was diagnosed with hypertension.

Analysis:

For diabetes, Welch's t-test was used to compare expenditures between early and late diagnosis groups. The test measured the statistical significance of the difference in mean expenditures and quantified the precision of this difference using a 95% confidence interval. Similarly, for hypertension, the same statistical approach was applied to identify any meaningful variations in expenditures between the two diagnosis groups.

Results:

For diabetes, a significant difference in expenditures was found, with a T-statistic of 4.976 and a p-value of 1.58×10^{3}. The mean expenditure for early-diagnosed individuals was $5,500 higher than for late-diagnosed individuals, with a 95% confidence interval of $3,333.886 to $7,666.590. In contrast, for hypertension, no significant difference was observed. The T-statistic was 1.438, the p-value was 0.151, and the mean difference was $1,611.530, with a 95% confidence interval of -$585.202 to $3,808.262, suggesting no substantial impact of diagnosis timing on expenditures.

Conclusion:

The findings suggest that earlier diabetes diagnoses are associated with significantly higher healthcare expenditures, reflecting the extended management of chronic conditions and complications. This highlights the importance of early intervention and preventive care to mitigate costs and improve outcomes. Conversely, the uniformity in hypertension-related expenditures suggests consistent treatment costs irrespective of diagnosis timing, necessitating further research to identify potential areas for care delivery improvement. These insights underscore the importance of addressing healthcare inequities and prioritizing early detection and support for chronic disease management.

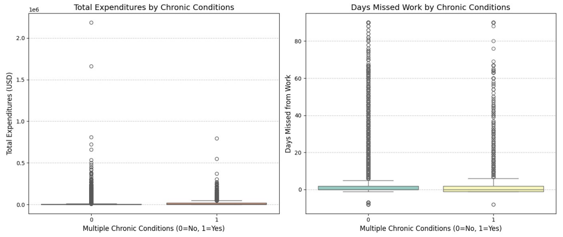
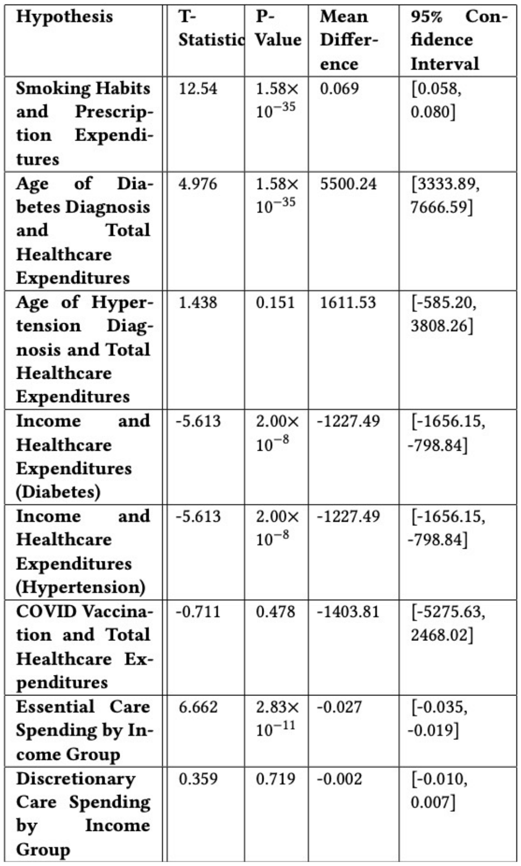


Figure 16: Relation Between Multiple Chronic Conditions and Healthcare Expenditures and Work Absence

5 HYPOTHESIS TESTING RESULT 2018-22:



Hypothesis	T-Statistic	P-Value	Mean Difference	95% Confidence Interval
Smoking Habits and Prescription Expenditures	12.54	1.58×10^{-35}	0.069	[0.058, 0.080]
Age of Diabetes Diagnosis and Total Healthcare Expenditures	4.976	1.58×10^{-35}	5500.24	[3333.89, 7666.59]
Age of Hypertension Diagnosis and Total Healthcare Expenditures	1.438	0.151	1611.53	[-585.20, 3808.26]
Income and Healthcare Expenditures (Diabetes)	-5.613	2.00×10^{-8}	-1227.49	[-1656.15, -798.84]
Income and Healthcare Expenditures (Hypertension)	-5.613	2.00×10^{-8}	-1227.49	[-1656.15, -798.84]
COVID Vaccination and Total Healthcare Expenditures	-0.711	0.478	-1403.81	[-5275.63, 2468.02]
Essential Care Spending by Income Group	6.662	2.83×10^{-11}	-0.027	[-0.035, -0.019]
Discretionary Care Spending by Income Group	0.359	0.719	-0.002	[-0.010, 0.007]

6 CONCLUSION

Our study delved into the intricate factors influencing healthcare expenditures, yielding both anticipated and unexpected insights. As expected, daily smokers incurred higher prescription drug costs, reflecting the well-known health risks associated with smoking. Similarly, individuals diagnosed with diabetes at a younger age faced greater long-term healthcare expenses, underscoring the importance of early intervention.

Contrary to our expectations, the timing of hypertension diagnosis did not significantly affect healthcare spending, suggesting consistent treatment approaches regardless of when the condition is identified. Additionally, COVID-19 vaccination status appeared to have minimal impact on overall healthcare costs, indicating that vaccination alone may not substantially alter expenditure patterns.

A particularly concerning finding was the disproportionate financial burden on low-income individuals with chronic conditions, highlighting systemic inequities in healthcare access and affordability. This underscores the urgent need for policies aimed at reducing disparities and ensuring equitable healthcare for all.

In summary, while some results aligned with our hypotheses, others revealed unexpected patterns, emphasizing the complexity of healthcare dynamics and the necessity for comprehensive strategies to address these multifaceted challenges.

7 FUTURE WORK

Several avenues for future research emerge from our findings:

Deeper Dive into Hypertension: A more in-depth analysis of hypertension-related factors, such as treatment adherence, comorbidities, and socioeconomic status, could provide valuable insights into the observed uniformity in healthcare expenditures.

COVID-19 Vaccination and Long-Term Health: Longitudinal studies are needed to assess the long-term health implications of COVID-19 infection and vaccination, particularly in vulnerable populations.

Healthcare Disparities and Policy Interventions: Further research is required to identify and evaluate effective policy interventions to address healthcare disparities, particularly for low-income individuals with chronic conditions.

Machine Learning Applications: Advanced machine learning techniques can be employed to develop predictive models for healthcare costs, enabling more targeted interventions and resource allocation.

By addressing these areas, we can gain a more comprehensive understanding of healthcare expenditures and inform evidence-based policies to improve health outcomes.

7.1 Data Sources

This study utilizes publicly available datasets from the Medical Expenditure Panel Survey (MEPS), provided by the Agency for Healthcare Research and Quality (AHRQ). The datasets, covering healthcare utilization, expenditures, and associated variables from 2018 to 2022, are accessible through the following links:

- **Data Source:** MEPS Data Files
- **2022 Data:** MEPS HC-243
- **2021 Data:** MEPS HC-233
- **2020 Data:** MEPS HC-224
- **2019 Data:** MEPS HC-216
- **2018 Data:** MEPS HC-209

The data files contain detailed information on healthcare expenditures, demographics, health status, and insurance coverage, serving as a robust foundation for the analysis conducted in this study.

hyperref

REFERENCES

[1] Kaiser Family Foundation. (2024). *Americans' Challenges with Health Care Costs*. Retrieved from https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/

[2] National Center for Biotechnology Information. (2024). *Health Care Costs: A Primer*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK557421/

[3] Smith, J. (2009). The cost of air medical transport. *Air Medical Journal*, 28(3), 125-129. Retrieved from https://www.airmedicaljournal.com/article/S1067-991X(09)00068-6/fulltext

[4] Harvard Communication Lab. (2023). Structure of a paper. Retrieved from https://communicate.gse.harvard.edu/files/commlab/files/$_{s}tructure_{o}f_{ap}aper.pdf$

[5] Paperpal. (2023). How to write a conclusion for research papers. Retrieved from https://paperpal.com/blog/researcher-resources/how-to-write-a-conclusion-for-research-papers