COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

FINAL PROJECT

BRFSS

DEPARTAMENTO DE TELEMÁTICA Y COMPUTACIÓN

ELENA CONDERANA MEDEM y SERGIO CUENCA NÚÑEZ

Índice

1. Introduction

Although having access to healthcare is essential for both individual and community well-being, many Americans struggle to do so because of financial constraints. Even though more people have health insurance now than ten years ago, many people state being unable to attend a doctor when needed due to the cost barrier it represents.

In this context, this report explores the question: **Which demographic groups are most affected by cost barriers to healthcare?** To address this, we have analyzed data from the Behavioral Risk Factor Surveillance System (BRFSS) in search of the self-reported cases of financial barriers to healthcare. The main goal is to identify trends and patters in income, education, work status, age, and sex to determine which groups have the most difficulty paying for necessary medical care.

By combining descriptive statistics and confidence intervals, we aim to move beyond general assumptions and provide statistical-based evidence insights into the inequalities that shape access to healthcare. This approach not only identifies areas where disparities exist but also provides a better understanding of how they differ among different segments of the population.

On the other hand, understanding the health behaviors of vulnerable groups of the population is also relevant. Therefore, this report further investigates a potential correlation between binge drinking and veteran status, investigating the question: **Are veterans more prone to binge drinking than the average population?** Veterans often face unique challenges upon returning home, including difficulties with employment, trauma, and mental health adjustments. These service-related experiences may contribute to higher rates of risky behaviors such as binge drinking. Binge drinking, as defined in the BRFSS dataset, is characterized by the consumption of 5 or more alcoholic beverages for men, or 4 or more for women, on a single occasion within the past 30 days.

To explore this potential association, the use of hypothesis testing will determine whether there is a statistically significant difference in the prevalence of binge drinking between the veteran population and the general population. By analyzing relevant data, the report aims to shed light on potential areas where targeted support and resources may be particularly important for the veteran community to mitigate future healthcare needs.

2. Data Sampling

2.1. Which demographic groups are most affected by cost barriers to healthcare?

This question requires identifying key demographic and socioeconomic factors that influence healthcare access. These are the variables that were chosen because of their relevance:

- `medical_cost_barrier`: Target variable, represents whether individuals could not access healthcare due to cost. It is a key public health indicator.
- `health_insurance`: Related to financial access to healthcare, uninsured individuals may be more likely to face cost-related barriers.
- `personal_doctor`: Reflects continuity of care and access to regular health services, which may influence both perceived and actual barriers.
- `sex`: A standard demographic variable included to explore if gender-based disparities exist.
- `age_group`: Age affects healthcare needs, insurance coverage, and employment status.
- `state`: Healthcare access and public assistance programs vary significantly by state.
- `income_level`: Lower income is directly associated with reduced access to care and increased cost sensitivity.
- `education_level`: Used to measure socioeconomic status, which can influence healthcare access.
- `employment_status`: Linked to both income and health insurance availability, especially in employer-based healthcare systems.
- `general_health`: Captures an individual's self-perceived health status, which may be correlated with healthcare utilization.
- `physically_unwell_days`, `mentally_unwell_days`, `activity_limited_days`: These variables measure recent health burden and its impact on daily life, potentially influenced by cost barriers.

For the purpose of making valid inferences about the broader U.S. population based on the BRFSS 2013 dataset, a representative sample had to be carefully selected. Instead of relying on simple random sampling, which could over or underrepresent important subgroups, we chose a stratified sampling strategy.

Specifically, we stratified the data according to income level and employment status, two important demographic factors that are closely linked to healthcare cost barriers. This reduces sampling bias and increases the accuracy of our results by guaranteeing that the sample fairly represents the range of income and sex categories seen in the whole dataset.

However, before sampling, the dataset needed to be cleaned so as missing values could be treated correctly. We eliminated records that lacked values for important characteristics such income, sex, and medical cost barrier status. This was a crucial step to maintain the integrity of the stratification and the validity of subsequent statistical analyses.

Using stratified random sampling, we selected a 10% subset of the cleaned dataset, proportionally preserving the distribution of income levels and sex. To validate that the sample was indeed representative, we compared the category distributions between the full cleaned dataset and the sample.

```
Full dataset — Income distribution:
 income_level
$75,000 or more      0.276353
Less than $10,000    0.060181
```

```
Less than $15,000    0.063477
Less than $20,000    0.082732
Less than $25,000    0.099139
Less than $35,000    0.116187
Less than $50,000    0.146486
Less than $75,000    0.155445
Name: proportion, dtype: float64
Sample — Income distribution:
 income_level
$75,000 or more      0.276324
Less than $10,000    0.060141
Less than $15,000    0.063488
Less than $20,000    0.082730
Less than $25,000    0.099128
Less than $35,000    0.116242
Less than $50,000    0.146456
Less than $75,000    0.155492
Name: proportion, dtype: float64

Full dataset — Sex distribution:
 sex
Female    0.57618
Male      0.42382
Name: proportion, dtype: float64
Sample — Sex distribution:
 sex
Female    0.574352
Male      0.425648
Name: proportion, dtype: float64
```

Listing 1. Full Dataset and Sample Distributions for CI Analysis

The Listing 1 confirms that the distributions matched nearly exactly, with absolute variations in proportions of less than 0.001 for every category. This demonstrates that the results yield by the EDA and confidence interval calculation in the following sections can be extrapolated to the larger U.S. adult population assessed by BRFSS in 2013, since the stratified sampling strategy effectively maintained the population structure.

2.2. Are veterans more prone to binge drinking than the average population?

To explore the research, question the following variables were chosen because of their perceived relevance:

- `vet`: Indicates whether the sampled person is a veteran or not.
- `sex`: A standard demographic variable included to explore the gender distribution among both groups.
- `age_group`: A standard demographic variable included to explore the age group distribution among both groups.
- `avgDrinksDay`: Measures during the past 30 days, on the days in which was drank, the average amount of drinks taken.
- `employment`: Specifies the current employment status of the interviewed.
- `lifeSatisfaction`: In a range between 'Very satisfied' and 'Very dissatisfied' it measures the overall satisfaction with their life. This variable has further implications in issues such as mental health or depression.
- `bingeDrinking`: The variable under study, which considering all types of alcoholic beverages

accounts for the amount of times a male surpassed 5 and a female 4 drinks in an occasion in the past 30 days.

Only 12,51% of the participants in the BRFSS dataset have veteran status. Therefore, to provide a more robust and representative comparison of binge drinking among veteran and non-veteran populations in the dataset, a stratified random sampling technique is proposed. This method ensures that both groups are adequately represented in the final sample, reflecting their proportions within the overall population. A simple random sample could underrepresent the veteran group, leading to a less reliable comparison.

Prior to sampling, the dataset underwent a cleaning process to address missing values and avoid idle records in the final sample. However, since the variables `avgDrinksDay` and `lifeSatisfaction` are relevant for exploratory data analysis and present a significant proportion (over 50% and 97% respectively) of NAs, missing value removal was selectively applied only to variables `vet`, `sex`, `age`, `employment` and `bingeDrinking`. Given the nature of the dataset, this decision was made balancing data completeness and the potential contextual insights offered by the variables with higher missing rates of data.

Following the cleaning stage, a 20% stratified random sample was generated, ensuring the proportional representation of the veteran population. The decision to use a larger sample size than in the confidence interval estimation was driven by the desire to retain as much information as possible from the variables with substantial missing data. To validate that the sample was indeed representative, we compared the category distributions between the full cleaned dataset and the sample.

```
Full dataset — Veteran distribution:
No    0.874863
Yes      0.125137
Name: veteran3, dtype: float64
Sample — Veteran distribution:
Yes    0.875074
No      0.124926
Name: vet, dtype: float64
```

Listing 2. Full Dataset and Sample Distributions for Hypothesis Testing

3. Exploratory Data Analysis (EDA)

3.1. Which demographic groups are most affected by cost barriers to healthcare?

Before making graphs and exploring the data, it is essential to check for biases that may be present in the data, and therefore, in the stratified sample that was chosen from it. Identifying and addressing biases ensures that the analysis and conclusions drawn are representative and reliable.

3.1.1. Self-Report Bias – General Health Distribution

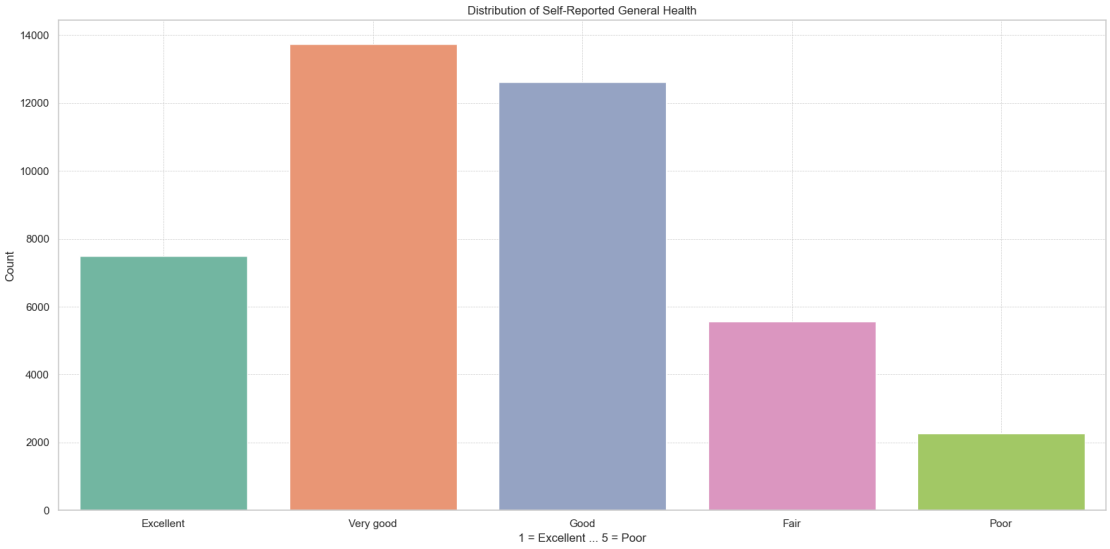

Figure 1. Distribution of Self-Reported General Health

Due to its complete reliance on self-reported data, BRFSS introduces self-report bias. People may not report unhealthy behaviors, inflate positive ones, or forget what happened to them. Specially, when it comes to controversial or delicate topics like money, healthcare access, or mental health, this is especially relevant.

We looked at the self-reported general health distribution to investigate this. Figure 1 shows that, with the majority of respondents choosing "Very Good" or "Good" health, there is a significant bias toward positive answers. This skew implies a social desirability bias, in which people present themselves in more favorably way than they actually are.

3.1.2. Coverage Bias – Top/Bottom States by Response %

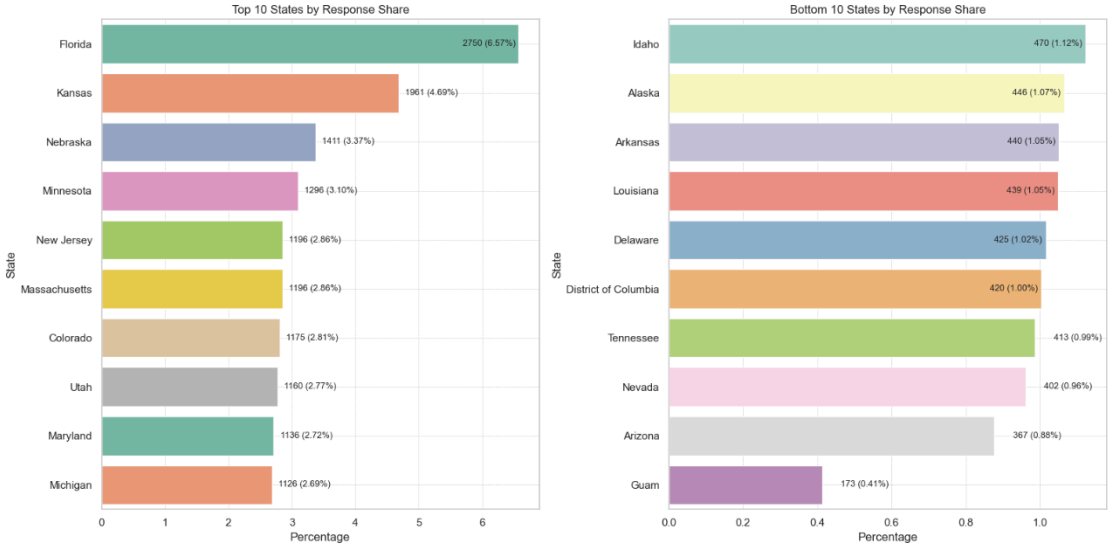

Figure 2. States by Response Rate

The BRFSS is a telephone survey that is intended for individuals in the United States who are not institutionalized, that means that certain vulnerable groups, such as those in prison, living in nursing homes, or without dependable phone service, are not included in the dataset.

We looked at the dataset's state distribution to investigate possible coverage bias. While some states, like Florida, have a 6.57% representation, others, including Tennessee, Nevada and Arizona, have fewer than 1%. Even though the discrepancies illustrated in Figure 2 could be the result of sample quotas or population size, some areas or groups could be underrepresented, having an impact on national estimates of healthcare access and cost barriers.

3.1.3. Missing Data Bias - Nonresponse by Variable

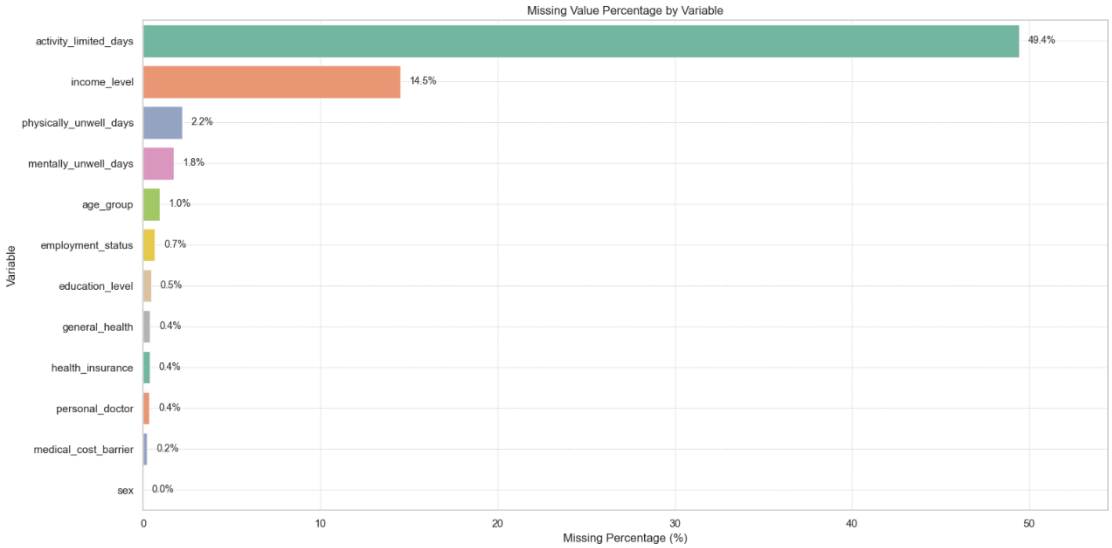

Figure 3. Missing Value % by Selected Variables

Lastly, another potential source of bias is missing data. Figure 3 depicts how the variable `activity_limited_days` is missing in around 49% of the cases, a significant source of possible bias in our dataset. The survey's design, including alternative modules and follow-up reasoning, may be the reason why there is a high nonresponse rate.

Furthermore, `income_level` is absent for 14.5% of respondents, which is particularly concerning because income is a crucial factor in determining access to healthcare. We eliminated missing income instances prior to stratified sampling and confidence interval calculation to reduce bias. However, if the missingness is not random, this might introduce systematic bias and underrepresent some vulnerable groups.

Only when those variables were directly engaged in a given analysis were rows dropped; other variables, such as `mental_health` and `employment_status`, had low levels of missing data and were handled using available-case analysis.

3.1.4. Impact of Cost Barriers on Health Metrics Across Selected Demographic Variables

To assess the impact of cost-related access barriers on health, we examined self-reported physically and mentally unwell days, along with activity-limited days, across income levels, cost barriers, and employment status.

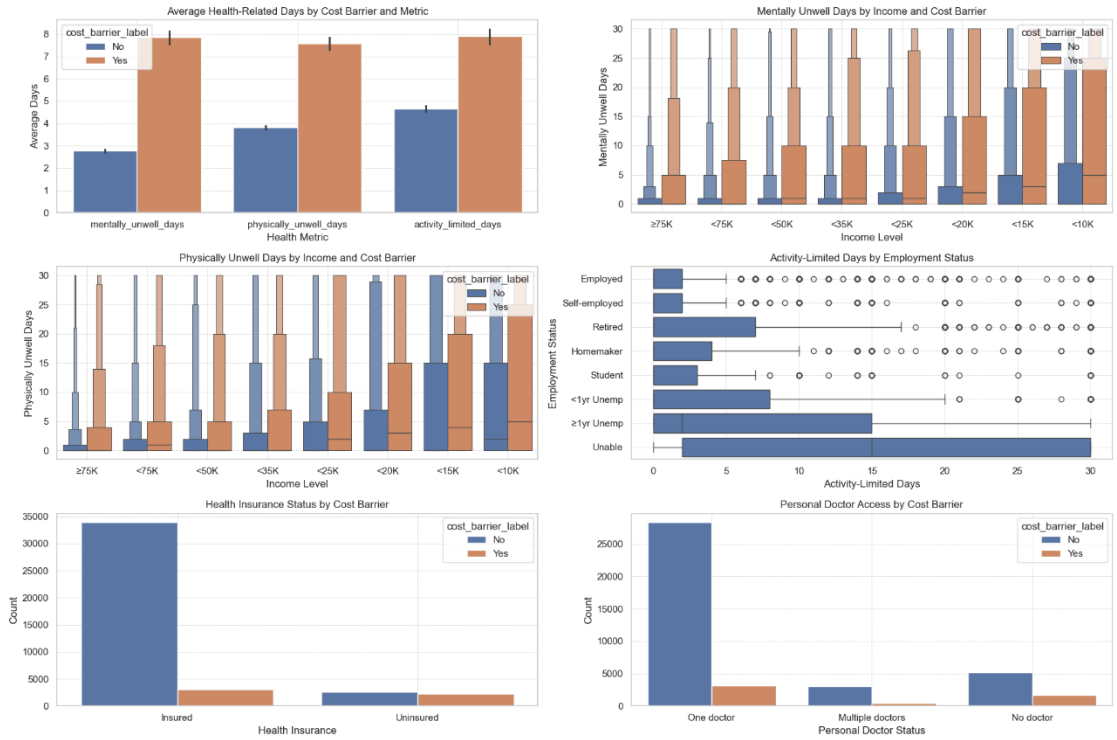

Figure 4. Impact of Cost Barriers on Health Metrics Across Demographics

These are the key findings and insights from the plots shown in Figure 4:

- Average Health Days by Cost Barrier and Metric: Individuals facing cost barriers frequently report more days with restricted activity, mental and/or physical illness compared to those without. The biggest disparity is in mental health.
- Mentally Unwell Days by Income and Cost Barrier: People with cost barriers report a much higher number of mentally unwell days than those in lower income categories, particularly those earning less than $10,000 and less than $15,000. The mental health disparity across cost barrier groups closes as income rises, indicating that psychological pressure may be mitigated by financial well-being.
- Physically Unwell Days by Income and Cost Barrier: As with mental health, those with lower incomes who report cost barriers are more likely to have physically ill days. The disparity is particularly noticeable in the <10K group and progressively narrows as income levels rise, suggesting that poverty and healthcare inaccessibility have a compounding effect.
- Activity-Limited Days by Employment: The largest percentage of activity-limited days are reported by those who are unemployed or unable to work. The relationship between job competence and perceived physical capability is further supported by the fact that employed people have the fewest functional limits.
- Health insurance Status by Cost Barrier: The majority of those covered by health insurance did not indicate a financial barrier. However, the percentage of uninsured persons who report a cost barrier is about the same as or greater than that of those who did not, indicating that not having insurance is a significant risk factor for running into problems with healthcare access due to costs.
- Personal Doctor Access: The number of people with multiple doctors is lower overall, with cost barriers being minimal in this group, possibly reflecting a wealthier or chronically ill subgroup with more access. Compared to those who have one or more doctors, individuals who do not have one are significantly more likely to report a cost barrier.

3.2. Are veterans more prone to binge drinking than the average population?

To explore potential differences in drinking behavior between veterans and non-veterans and assess whether veterans show a greater tendency toward binge drinking, we examined several demographic and behavioral indicators. These included gender and age distribution, employment status, average alcohol consumption, and self-reported life satisfaction. In the age-gender alcohol consumption analysis, extreme values, where average drinks per day exceeded 10, were excluded, as they were considered outliers and even faulty points, having values up to 76, that clouded the overall pattern.

Figure 5. A Comparison of Employment Profiles, Drinking Patterns, Demographic Attributes, and Life Satisfaction among Veterans and Non-Veterans.

These are the key findings and insights from the plots shown in Figure 5:

- Employment Status Distribution: Veterans and non-veterans showcase significantly different employment patterns. Retirement constitutes the primary employment status for over half of the veteran population, a stark contrast to non-veterans who are predominantly employed or pursuing education. These imbalances in labor force distribution are likely influenced by the different age profiles and the limited participation of younger veterans in the dataset, which are also scarcer.

- Alcohol Consumption by Employment Status: Drinking behaviors differ substantially across employment statuses. Students and individuals who are unemployed, especially those out of work short-term, report the highest average alcohol consumption per day. This trend may reflect the influence of social isolation or psychological stress. In contrast, retired

individuals exhibit relatively lower average alcohol intake, suggesting that alcohol use is shaped more by personal and socioeconomic instability than simply by availability of free time.

- Demographic Composition by Age and Veteran Status: The age distribution of veterans differs substantially from that of non-veterans, with a notable concentration in the ages of 55 and older. In contrast, non-veterans are more evenly distributed across the different age groups, with their highest representation in the 55 to 59 age bracket.
- Gender Representation among Veterans: Veterans are predominantly male, while the non-veteran population shows a more balanced gender distribution, skewing slightly towards females. Since men are usually contemplated as more prone to higher alcohol consumption, the male prominence might suggest a predisposition toward heavier drinking among veterans.
- Drinking Trends across Age and Gender: Heavier alcohol consumption, including binge drinking, is concentrated among younger age groups, especially males under the age of 35. Both average daily intake and binge drinking proportions decline sharply with age. While veterans are predominantly male, a group that displays higher drinking levels across nearly all age categories, this is counterbalanced by their older age profile. Since veterans are largely concentrated in older age brackets, where alcohol use is substantially lower, age appears to offset the expected impact of gender. This dynamic suggests that the veteran population may not engage in higher alcohol consumption than non-veterans, despite being mainly composed of men, who appear to have a greater drinking disposition.
- Life Satisfaction Patterns: The overall distribution of life satisfaction showcases a very similar appearance between veterans and non-veterans with no extreme differences suggesting widespread depression or significant dissatisfaction among veterans. This similarity may indicate that, despite the challenges associated with military service and reintegration, veterans maintain levels of life satisfaction comparable to the general population, which could reflect resilience or effective support systems post-service.

4. Inference Statistics

4.1. Which demographic groups are most affected by cost barriers to healthcare?

Before performing the confidence interval tests, we needed to determine which demographic variables were most relevant for analyzing cost-related barriers to healthcare. To achieve that, we used two standard association measures:

- Cramér's V: To assess the strength of association between categorical variables and the target variable (`medical_cost_barrier`).
- Eta squared (η^2): To evaluate the proportion of variance explained by the target variable in continuous health-related measures.

Variables with higher values of Cramér's V were selected for proportion and confidence interval analysis. These included `health_insurance`, `income_level`, `employment_status`, `general_health`, `age`, and `personal_doctor`, as they showed the strongest relationships with the presence of cost barriers. On the other hand, for continuous variables, those with higher eta squared values, specifically `mentally_unwell_days`, `physically_unwell_days`, and `activity_limited_days`, were selected for t-tests. These variables reflect individuals' physical and mental health burden and showed significant differences based on whether respondents reported a cost barrier.

The relationship between cost barriers and key categorical variables was analyzed using proportions and 95% confidence intervals. The results are shown in Figure 5.

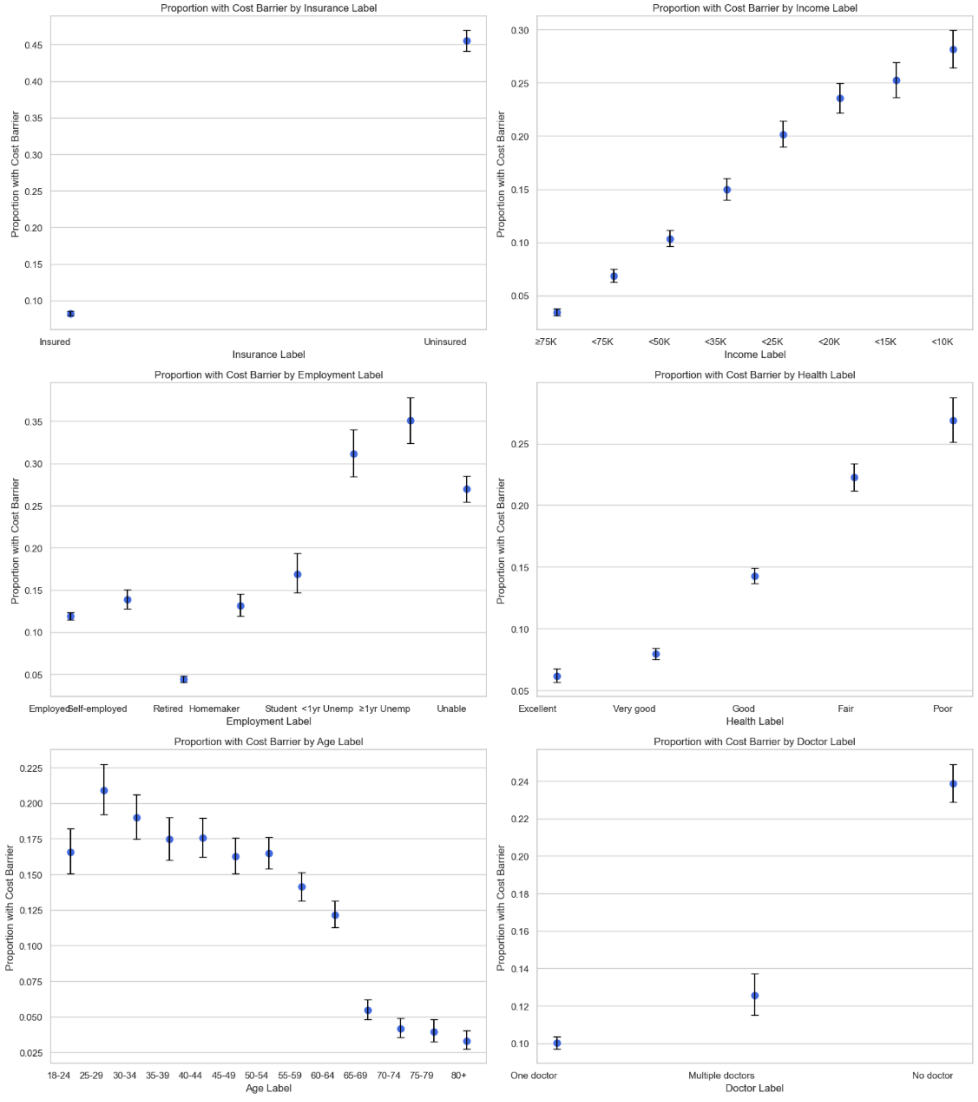

Figure 5. Categorial Variables' Confidence Intervals for Medical Cost Barriers

Health Insurance Status

The strongest predictor of medical cost barriers was the presence or absence of health insurance. Among uninsured individuals, 45.5% reported experiencing a cost-related barrier (95% CI: 44.1%–47.0%), compared to just 8.2% of those with insurance.

Income Level

There is an inverse relationship between income and cost barriers. For those earning $75,000 or more, only 3.4% reported barriers, while that figure rose to 20.2% for individuals earning under $25,000. Among those with incomes under $10,000, 28.1% faced barriers (95% CI: 26.4%–29.9%).

Employment Status

Unemployed individuals were also far more likely to encounter a pricing barrier. The barrier rate was 35.1% for those who had been without a job for more than a year (95% CI: 32.4%–37.8%). and 31.2% for those who had been unemployed for less than a year (95% CI: 28.4%–34%). In contrast, just 11.9% of those in employment mentioned problems linked to costs. At 26.9%, those who were unable to work because of a disability or sickness also faced significant obstacles.

General Health Status

Cost barriers were strongly associated with self-perceived health status. Only 6.2% of respondents who rated their health as "excellent" reported medical cost barriers (95% CI: 5.66%–6.75%), whereas 26.9% of those in "poor" health did (95% CI: 25.42%–28.55%).

Age Group

Moreover, younger adults were more likely to face cost barriers. For those aged between 18-24, 16.6% reported cost barriers (95% CI: 15.05%–18.21%), compared to just 3.3% among individuals aged 80 and older (95% CI: 2.73%–4.02%). This clearly reflect lower coverage rates or job instability among younger adults.

Access to a Personal Doctor

From those surveyed, individuals that do not have a regular doctor reported higher cost barriers (23.9%) (95% CI: 22.88%–24.91%) than those with one doctor (10.0%) (95% CI: 9.67%–10.35%) or multiple doctors (12.6%) (95% CI: 11.51%–13.73%). The absence of continuous care seems to correlate with financial barriers to accessing the healthcare system.

Three continuous variables were analyzed using boxplots and t-tests to examine differences in health outcomes between individuals with and without cost barriers.

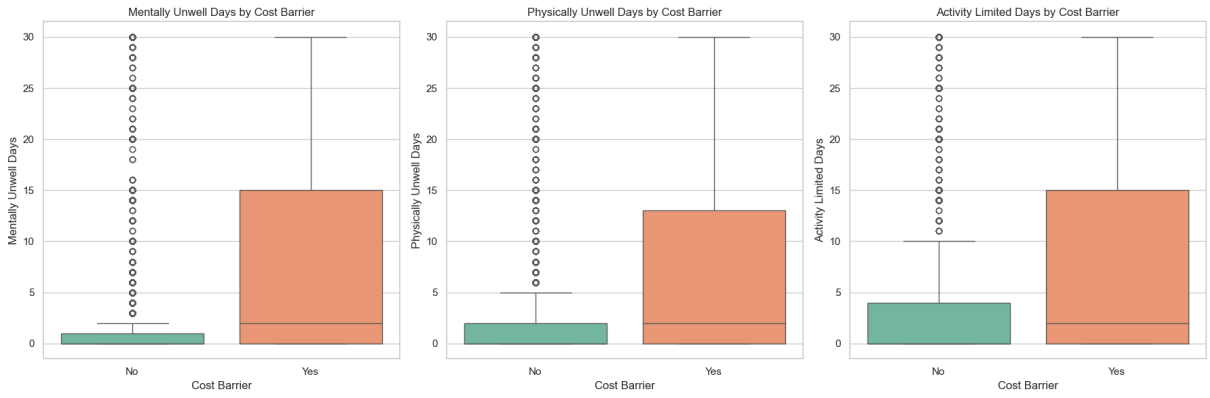

Figure 5. Continuous Variables T-Tests for Medical Cost Barriers

Mentally Unwell Days

Those who reported a financial barrier had an average of 7.87 days of mental illness during the

previous month, whereas those who did not had an average of 2.77 days. This implies that people with barriers have more than twice as many days with mental health problems, suggesting that financial access and mental health are closely related.

Physically Unwell Days

Moreover, the physical health impact was significantly higher for individuals who faced financial restrictions, with an average of 7.83 days of physical sickness compared to 3.74 days for those who did not. Again, it is also duplicated, with clearly differentiated distributions.

Activity-Limited Days

A similar pattern was observed in days where respondents were limited in their usual activities due to health issues. On average, people with difficulties reported 7.74 days, whereas people without barriers reported 4.58 days. This leads to conclude that cost barriers are associated with poorer overall quality of life.

4.2. Are veterans more prone to binge drinking than the average population?

After extracting some insight from the underlying data, the hypothesis testing will determine whether veterans exhibit a higher proportion of binge drinking compared to non-veterans. The statistical hypotheses for this test are as follows.

$$\boldsymbol{H_0}: p_{non-vet} \geq p_{vet}$$

$$\boldsymbol{H_A}: p_{vet} > p_{non-vet}$$

The null hypothesis (H_0) states that the proportion of binge drinking among veterans is less than or equal to that of non-veterans, whereas the alternative hypothesis (H_A) asserts that the proportion is higher among veterans. To answer whether veterans are more prone to binge drinking a one-sided, right-tailed test is needed to compare the proportions. The focus lies not exclusively in whether there is any difference between both groups, but in whether veterans binge drink more.

The resulting scores of the hypothesis testing were a **z-score of -1.433** and a **p-value of 0.0759**. Since the p-value is greater than the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Furthermore, the negative sign of the z-score reinforces the idea that, in this sample, veterans may have a lower proportion of binge drinking compared to non-veterans. Figure 6 summarizes these findings comparing the binge drinking proportions between both populations.

The bar chart displays the proportion of binge drinkers in each group with confidence intervals illustrating the range within the true population is likely to lie. Non-veterans show a barely higher proportion, which alongside the overlap in the confidence Intervals suggest that the true proportions in the populations may not be significantly different. Hence, the statistical evidence does not support the claim that veterans are more likely to engage in binge drinking.

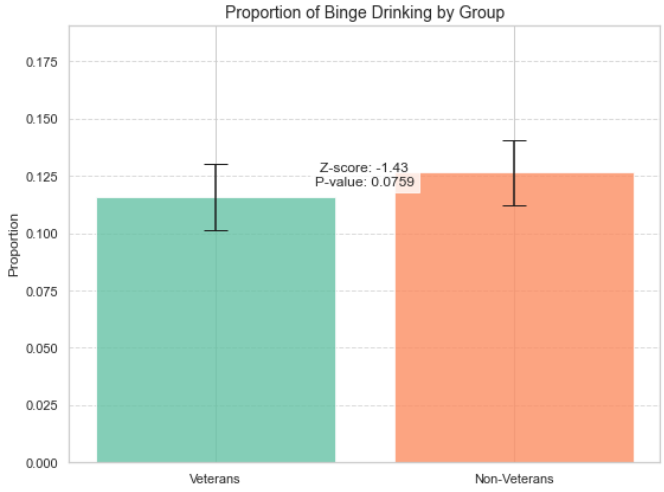

Figure 6. Proportion of Binge Drinking between Veterans and Non-Veterans along with 95% CI

5. Conclusions

5.1. Which demographic groups are most affected by cost barriers to healthcare?

The results clearly indicate that cost barriers in access to healthcare disproportionately affect:

- People without health insurance.
- Individuals with low income.
- Unemployed or unable to work.
- In poorer health.
- Without regular access to a doctor.
- And to a lesser extent, young people.

Additionally, individuals facing cost barriers tend to experience more days of mental distress, physical illness, and limitations in daily activity. These results suggest a direct connection between financial constraints and both physical and mental health outcomes, underscoring the importance of addressing cost-related access issues to improve population health. Table 1 reveals the main findings of the carried-out research.

Finding	Explanation
Uninsured individuals → more barriers	Lack of insurance limits access due to out-of-pocket costs.
Low income → more barriers	Relative cost is higher; financial constraints reduce access.
Unemployment or exclusion from labor market → more barriers	Less access to employer-based insurance and income instability.
Poor general health → more barriers	Greater healthcare needs lead to higher potential costs.
Younger age → more barriers	Often lack coverage, face job instability or temporary work.
More physical and mental distress among those with barriers	Limited access likely contributes to worse health outcomes.

Table 1. Demographic Groups Affected by Medical Cost Barriers

5.2. Are veterans more prone to binge drinking than the average population?

In conclusion, the statistical evidence from this analysis does not support the claim that veterans are more likely to engage in binge drinking. While the observed trend in our sample even suggested a slightly lower proportion of binge drinking among veterans compared to non-veterans, though not statistically significant, it is crucial to avoid drawing definitive statements about the broader populations based solely on this analysis. The outcome of this analysis, however, underscores the importance of grounding assumptions in data rather than relying on preconceived notions.

The interpretation of these findings needs to account for a couple of key differences between our veteran and non-veteran samples. Veterans were more likely to be retired, which could have influenced their health behaviors, including alcohol consumption. Additionally, the older age of the veteran group might have affected the overall rates of binge drinking found, as this behavior varies with age as addressed. Further analysis that delves deeper into these specific subgroups, for example comparing employed veterans to employed non-veterans, or examining binge drinking rates within narrower age brackets, could potentially yield more specific insights.

Furthermore, inherent limitations of the BRFSS dataset must be acknowledged. The reliance on self-reported data introduces potential biases related to recall and social desirability with variables

such as `lifeSatisfaction`. Additionally, the presence of underrepresented groups and substantial missing data could affect the generalizability and accuracy of our findings. These factors highlight the need for further investigation using larger, more representative samples to gain a deeper understanding of alcohol consumption patterns within the veteran community.