CPSC 368 Project Proposal

KNM Neighbours (Nicholas Tam (45695970), Kevin Shiao (73239121),

Minghao Wang (56536469))

1. Introduction

Modern technologies have improved our quality of life, making everything more convenient than ever by reducing workload and stress from various sources. This allows people to focus on their health more efficiently. With modern health information technologies, users can receive immediate feedback on their physical condition anytime, anywhere (Li et al., 2019). Health insurance is important and beneficial, as uninsured individuals often experience poorer health and receive less medical care, often with delays (Bovbjerg, Hadley, 2006). However, the extent to which health insurance impacts health remains debatable. As Levy and Meltzer suggest, determining whether health insurance plays a significant role in influencing health will likely require substantial investment in social experiments (Levy, Meltzer, 2008). Although this paper does not directly answer the ultimate question or fill the gap, it contributes by presenting significant findings that serve as supporting evidence, aiming to attract attention in the healthcare and health insurance fields. Specifically, we explore how health insurance coverage impacts health outcomes among U.S. adults. The impact of health insurance will be measured in three ways: (1) by sex (male and female), focusing on coronary heart disease mortality by sex; (2) by state, examining coronary heart disease mortality across different states; and (3) by disease, comparing coronary heart disease mortality with various cancer mortalities.

2. Methodology

The datasets ‘U.S. Chronic Disease Indicators’ and ‘Health Insurance Coverage of Adults Ages 19-64’ are from HealthData.gov and KFF respectively. The ‘U.S. Chronic Disease Indicators’’ dataset contains 4820 observations and the attributes used are YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, and Stratification1 (Centers for Disease Control and Prevention, 2024). On the other hand, the datasets ‘Health Insurance Coverage of Adults Ages 19-64’, ‘Health Insurance Coverage of Women Ages 19-64’, and ‘Health Insurance Coverage of Men Ages 19-64’ from KFF each contain 52 observations, one for each state in the US, and the attributes used are Location and Uninsured (KFF, 2024). Data exploration, evaluation, and cleaning will be conducted using Python. To more effectively explore the datasets, the two tables will be joined by state. Once these processes are completed, the cleaned data will be structured and stored in an SQL database, with key questions addressed using SQL queries.

Since the combined table does not include population percentages for males and females, additional methods are needed to approximate insurance distribution by sex. To measure the impact by sex, total population data and male/female population percentages for the examined states will be obtained from government-authorized websites to ensure accuracy and reliability. From there, the percentage split between males and females in the examined states is determined by:

$$\%_{sex} * N_{total} = N_{sex}$$

$$PI_{x} * N_{total} = N_{x}$$

$$PI_{sex_{x}} = PI_{x} * \left(\frac{N_{sex}}{N_{total}}\right)$$

Where $\%_{sex}$ is the population percentage of sex (male or female), N_{total} is the total population, N_{sex} is the population of sex (male or female), PI_x is the proportion insured by x (insured or uninsured), N_x is the proportion of the population insured by x (insured or uninsured), and PI_{sex_x} is the proportion of sex (male or female) insured by x (insured or uninsured).

2.1 Data Evaluation

Before each evaluation, QQ plots and residual plots will be created to validate model assumptions and ensure robustness and accuracy. Specifically, we will check the normality of residuals, homoscedasticity, and independence. If any assumptions are violated, appropriate corrective measures, such as data transformations, will be applied.

Impact by Sex

For this analysis, we will use Multiple Linear Regression (MLR) to examine the relationship between uninsured rates and coronary heart disease (CHD) mortality rates across males and females in Texas and Massachusetts. The model will include uninsured rates as a continuous predictor and sex as a categorical variable. Additionally, we will incorporate an interaction term between the uninsured rate and sex. This will allow us to determine if higher uninsured rates are more strongly associated with CHD mortality in one sex compared to the other. By employing MLR with interaction, we allow for a deeper understanding of relationships between variables by capturing the potential effects between variables, making our approach more scientifically robust. For more effective interpretations, we will also visualize our model to display how the relationship between uninsured rates and CHD mortality varies by sex.

Impact by State

We will also investigate the relationship between uninsured rates and coronary heart disease (CHD) mortality across all US states. First, we will gather data on the uninsured rates and CHD mortality rates for each state. Then, we will use Support Vector Regression (SVR) to analyze mortality rates and uninsured rates across the US to see if there are any notable relationships between them. We chose to apply SVR because it can capture the non-linear trends in the data that other models may not be able to capture. Additionally, we will perform hyperparameter optimization to fine-tune our model and ensure it is as accurate as possible.

Impact by Disease

Lastly, to compare the impact health insurance has on different diseases, we will develop two separate Poisson regression models: one to investigate the relationship between uninsured rates and coronary heart disease mortality rates, and another to examine the same relationship for cancer mortality rates in Texas and Massachusetts. By looking at each disease independently, we can ensure that the models are fitted to the data distribution without any interaction between the two diseases. We chose to use Poisson regression because it is specifically used to model count data. Poisson regression models the rate of an event's occurrence and provides rate ratios that indicate how changes in uninsured rates influence the likelihood of mortality. After fitting both models, we will compare the results to assess whether the effect of uninsured rates differs between cancer and CHD mortality rates. This comparison will help us understand if the uninsured rate has a stronger or weaker impact on mortality for one disease relative to the other, providing valuable insights.

2.2 Data Trustworthiness

The US Chronic Disease Indicators dataset (Centers for Disease Control and Prevention., 2024) is sourced from the Centers for Disease Control and Prevention, the national public health agency of the United States and a federal agency under the Department of Health and Human Services (HHS Office of the Secretary and Office of Budget (OB), 2019). Each row consists of a location, a topic, a corresponding question, a data value unit (including raw numbers, cases per 10,000 people, and percentages), type and value, a stratification category, and a corresponding subcategory. The data collected is derived from the American Community Survey, state alcohol sales from the Alcohol Epidemiologic Data System, the American NonSmokers' Rights Foundation database, surveys from the Behavioral Risk Factor Surveillance System, administrative and claims data from Centers for Medicare & Medicaid Services, the Current Population Survey Food Security Supplement, the National Immunization Survey, the National Survey of Children's Health, census from the National Vital Statistics System, surveys from the Pregnancy Risk Assessment Monitoring System, the U.S. Cancer Statistics Data Visualizations Tool, administrative, claims data and other data from United States Renal Data System, and the Women, Infants, and Children Participant and Program Characteristics Study (Centers for Disease Control and Prevention, 2024).

The Health Insurance Coverage of the Total Population datasets for 2019 and 2021 (KFF, 2024) are sourced from the Kaiser Family Foundation (KFF), which has been praised for being the “most up-to-date and accurate information on health policy” (Wonkblog Team, 2013). According to the KFF website, “the majority of our health coverage topics are based on analysis of the Census Bureau’s American Community Survey (ACS) by KFF. ACS includes a 1% sample of the US population and allows for precise state-level estimates. The

ACS asks respondents about their health insurance coverage at the time of the survey. Respondents may report having more than one type of coverage; however, individuals are sorted into only one category of insurance coverage." The KFF estimates are derived from the American Community Surveys, with our dataset being obtained from the one in 2019 (KFF, 2024).

AI Tool Use Declaration

We have used Chegg from Cite This For Me to assist with citations and ChatGPT for grammar checking.

References

- Centers for Disease Control and Prevention. 2024. U.S. Chronic Disease Indicators. (March 2024). Retrieved February 9, 2025 from https://healthdata.gov/dataset/U-S-Chronic-Disease-Indicators/dhcp-wb3k/about_data
- KFF. 2024.(October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/adults-19-64/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- HHS Office of the Secretary and Office of Budget (OB). 2019. Centers for Disease Control and Prevention. (November 2019). Retrieved February 9, 2025 from https://web.archive.org/web/20200410150453/https://www.hhs.gov/about/budget/fy-2020-cdc-contingency-staffing-plan/index.html
- Wonkblog Team. 2013. Presenting the third annual WONKY Awards - The Washington Post. (December 2013). Retrieved February 9, 2025 from

https://www.washingtonpost.com/news/wonk/wp/2013/12/31/presenting-the-third-annual-wonky-awards/

- KFF. 2024.(October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/health-insurance-coverage-of-women-19-64/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- KFF. 2024.(October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/health-insurance-coverage-of-men-19-64/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- Centers for Disease Control and Prevention. 2024a. Indicator Data Sources. (June 2024). Retrieved February 10, 2025 from https://www.cdc.gov/cdi/about/indicator-data-sources.html
- Junde Li, Qi Ma, Alan HS. Chan, and S.S. Man. 2019. Health monitoring through wearable technologies for older adults: Smart wearables acceptance model. *Applied Ergonomics* 75 (February 2019), 162–169. DOI: http://dx.doi.org/10.1016/j.apergo.2018.10.006
- Randall R. Bovbjerg and J. Hadley, "Why Health Insurance Is Important," *Urban Institute*, 2006. [Online]. Available: https://www.urban.org/sites/default/files/publication/46826/411569-Why-Health-Insurance-Is-Important.PDF. [Accessed: 10-Feb-2025].
- Helen Levy and David Meltzer. 2008. The impact of health insurance on Health. *Annual Review of Public Health* 29, 1 (April 2008), 399–409. DOI: http://dx.doi.org/10.1146/annurev.publhealth.28.021406.144042