Group Proposal

Alicia Helfrich, Olivia Gomez, and Akbar Naqvi

Motivating Question

Our team will be seeking to explore the question "What percent of individuals should be eligible for Medicaid after COVID-19 but are not presently covered?". After the COVID-19 public health emergency states were required to reevaluate the eligibility requirements for Medicaid, resulting in the loss of Medicaid coverage for many individuals. Given that many individuals became vulnerable both in terms of economic and health status following the pandemic, it is theorized that a large quantity of those that lost coverage would still greatly benefit from these services.

Data Sources and Methods

We will be exploring this data through the utilization of the 2019 and 2022 US Census Community Resilience Estimates and the 2019 and 2022 US Census Survey of Income and Program Participation (SIPP). We will first conduct an exploratory data analysis of both datasets to identify high level trends, assess missingness, and identify potential categorical variables of interest that may need to be grouped/combined or dropped to reduce the impact of outliers.

The Community Resilience Estimates are available at a Census Tract, County, and State level. It provides estimates of a "Community Resilience Index" that assesses factors most correlated with the capacity of households to respond to, absorb, and recover from disasters. One factor included within the Index is the the percent of households that do not have health insurance. In order to target our investigation of Medicaid and health care coverage on an individual level, we will identify the top 15 states that had the greatest reduction in Community Resilience in terms of insurance coverage between 2019 and 2022. We believe that this will be a strong proxy indicator for states that changed their eligibility requirements following the pandemic, and therefore will be an effective way to streamline our analysis when examining data on an individual level. During the selection of states, we will utilize parametric and non-parametric methods to create a new variable of the true estimated percent change in insurance coverage between 2019 and 2022. We will select the states with the largest percent reduction in health insurance coverage as predicted by the new variable (either parametric or non-parametric) with the estimated smallest margin of error.

The 15 states identified in the above analysis will then be utilized to filter the responses of the SIPP data. In order to assess true eligibility of Medicaid versus actual coverage in 2022, we will create a supervised machine learning model to predict coverage. 2019's data will be utilized as the testing and training data to represent "true" eligibility for Medicaid coverage and select the model that is the best predictor of this variable. We will then utilize the 2022 dataset as the implementation data to evaluate whether individuals surveyed should actually be covered by Medicaid. These predicted values will be compared against whether an individual actually had Medicaid coverage in 2022, and then results will be aggregated at a state level to assess the percent of individuals that are estimated to be eligible for coverage but are not covered.

Anticipated Technical Hurdles

There are several potential technical hurdles we anticipate throughout the analysis:

- **Community Resilience Estimates may not be an effective indicator of changes to Medicaid Eligibility policies.** To mitigate, we will communicate the limitations of the analysis within our final report, in order to ensure proper interpretation of the results.
- **SIPP data only provides state level changes, when analysis may be more effective at a county or tract level.** Unfortunately, given restrictions around personal identifying information, SIPP data doesn't include more geographic details beyond states. This is an unavoidable limitation, so we will be sure to communicate how this may impact interpretibility within our final report.
- **Data missingness at an individual level within SIPP data.** Given that SIPP data was collected through a series of individual surveys, it will be likely that there will be patterns of missing data, as occurs with any surveys. We will assess the nature of missingness, and then act accordingly to ensure the integrity of the analysis or communicate limitation to generalizability. This could include dropping observations or imputing data through an appropriate method.
- **Given differences in state Medicaid laws and procedures, it will be difficult to scale this evaluation on a national level.** The analysis will be looking high-level at the impacts of changes on Medicaid laws. However, it will not reveal any specific changes in laws and procedures that most impacted eligibility and therefore, give the variance in policy changes on a state level, it will be hard to translate the analysis to other states and jurisdictions.
- **Potential changes in SIPP variables collected between 2019 and 2022.** In order to properly utilize a machine learning model developed from 2019 data and apply it to 2022 data, there will need to be close alignment between information being available between the two datasets. We will need to limit the development of the model for 2019 to variables that are also present in the 2022 dataset.