**HEALTHCARE DATA ANALYSIS – HOSPITALIZATION RISK PREDICTION:**

Assessing Healthcare Access and Predicting Hospitalization Risk Using Socioeconomic and Geographic Factors.

## Project Overview:

This proposal outlines a data analysis project aimed at examining healthcare accessibility and developing a predictive model for hospitalization risk across U.S. census tracts using social determinants of health (SDOH). By analyzing the relationships between healthcare access, socioeconomic status, and geographic factors, we seek to identify communities at heightened risk and provide actionable insights for public health officials, policymakers, and healthcare providers to address gaps in healthcare accessibility.

## 1 Introduction:

Healthcare accessibility remains a critical challenge in the United States, with considerable disparities in access and outcomes across different communities. Direct hospitalization data can be challenging to obtain, but social determinants of health—such as access to medical facilities, insurance coverage, poverty levels, disability rates, and other demographic or environmental factors—serve as valuable proxies for predicting hospitalization risk.

This project will analyze the Agency for Healthcare Research and Quality (AHRQ) Social Determinants of Health Database to understand these relationships and develop a predictive

model for hospitalization risk. We expect our findings to provide actionable insights that can help address health care access disparities.

**2 Problem Statement:**

Many communities face barriers to healthcare access, which can lead to delayed treatment, increased emergency visits, and higher hospitalization rates. However, hospitalization risk factors are often hidden within socioeconomic and geographic disparities.

This project will:

- Clean and preprocess raw SDOH data to identify missing values and inconsistencies.
- Perform hypothesis testing to analyze disparities in healthcare access.
- Develop a predictive model to estimate hospitalization risk based on socioeconomic and geographic factors.

**3 Target Audience:**

The intended audience for this project is a public health agency in the United States in one of the four regions identified by AHRQ, specifically Northeast, Midwest, South or West.

This project specifically aims to provide these agencies with actionable insights into how socioeconomic and geographic factors influence hospitalization risk at the census tract level. By identifying high-risk areas and the key drivers behind hospitalization patterns, public health officials can make data-informed decisions about where to direct interventions, improve access to care, and support vulnerable populations more effectively. Additionally, this work may also

serve as a decision-support tool for grant application prioritization, community health needs assessments, and strategic health planning at the regional level.

**4 Research Question and Hypothesis:**

This project will analyze a 2020 dataset from the U.S. and its territories to investigate the following research question using statistical analysis and predictive modeling:

**Question:** Can social determinants of health–such as access to medical facilities, insurance coverage, poverty levels, disability rates, and other demographic or environmental factors–serve as proxies for predicting hospitalization risk?

**Hypothesis**: Socioeconomic and geographical factors contribute to predicting hospitalization risk.

**Prediction:** A predictive model incorporating distance to healthcare facilities, poverty levels, health insurance coverage, disability status, and household size will effectively estimate hospitalization risk. We expect that greater distance to care, higher poverty, lack of insurance, having a disability, and a smaller household will be associated with increased hospitalization risk.

**Novelty**: Our research question is not entirely novel, as previous studies have examined the relationship between social determinant of health (SDOH) and hospitalization risk. Factors such as poverty, insurance coverage, and access to care have been well-documented in public health research as key contributors. However, we will be applying machine learning techniques to predict hospitalization risk using a comprehensive dataset at the census tract level which will give us more granular insights.

**5 Variable Justification and Intersections:**

The chosen predictor variables—distance to care, poverty level, and health insurance coverage—are grounded in public health research and the social determinants of health framework. These factors are not only significant on their own but also interact in meaningful ways. For instance, individuals in high-poverty neighborhoods are more likely to live farther from hospitals and have lower rates of health insurance coverage. This geographic and economic overlap can compound barriers to care, increasing the risk of hospitalization. Understanding and quantifying these intersections will be a key focus of our analysis.

**6 Data and Methods**

For this project, we use the Social Determinants of Health (SDOH) Database provided by the Agency for Healthcare Research and Quality (AHRQ). This dataset, sourced from the U.S. Census Tract data for 2020, includes 85,500 observations and 329 features spanning a wide range of socioeconomic and geographic factors. The dataset is a strong choice for our research question, as it provides comprehensive demographic, economic, and environmental indicators that can help assess hospitalization risk based on social and geographic determinants.

The dataset was acquired directly as a CSV file from AHRQ's website. No merging was necessary since this is a single, self-contained dataset. However, initial exploratory analysis revealed that 19 columns contained more than 30% missing values, which we decided to remove due to insufficient data coverage. The remaining 310 columns have less than 30% missing values and will be retained, though missing values will be handled through imputation techniques such as mean, median, or mode replacement. Further preprocessing steps include normalizing

numerical variables and converting categorical variables into numeric features for model compatibility. These steps ensure data quality and allow for effective machine learning modeling.

**Data Source**: Agency for Healthcare Research and Quality (AHRQ) - SDOH Database at https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html

**Link to codebook in Github:** https://github.com/lemieuxjm-cap/Iota-Capstone/reports/codebook.md

## 7 Analysis Plan

Our anticipated analysis plan includes using R and/or Python to perform regression-based machine learning. Throughout our work, we will be watching for overfitting. Below is a breakdown of the work we plan on performing.

**Pre-processing steps:** In addition to the data cleansing techniques mentioned in the previous section, we will:

- Normalize numerical variables, including income, distance and pollution levels
- Encode categorical variables, including region, state, county
- Identify training and test data sets in a 70% training and 30% testing ratio
- Set a seed value and initialize cross validation and metrics

**Initial Model:** Our unsupervised learning model will be a Principal Component Analysis (PCA) to achieve goals of dimension reduction, important feature identification, and pattern discovery.

This will be supplemented with results of a K-means clustering to identify the appropriate number of clusters for grouping. To determine the optimal number of clusters for K-Means, we will use the elbow method and Silhouette score analysis. These methods will help us identify the cluster separation and assess the quality of resulting clusters. We will know if our unsupervised PCA model is performing sufficiently if we can explain more than 85% of the variance with a reduced number of dimensions.

**Second Model:** Our supervised learning model will be a Lasso Regression model to establish a hospitalization risk from our now-reduced number of variables, using the training and test sets previously established. For hyperparameter tuning, we will perform a random search for optimal parameters and use cross-validation to improve generalization.

**Evaluation:** We will use some or all of the following methods to evaluate the performance of the regression model:

- Accuracy
- Confusion matrix and its subcomponents:
    - True and false positives
    - True and false negatives
    - Precision
    - Recall (sensitivity)
    - F1-Score
- ROC-AUC and related calculations
- RMSE/MAE

Our model will be considered performant if we minimize overfitting, have fairly high accuracy, an ROC calc with a large AUC, and a low RMSE/MAE.

**8 Planned Visualization:**

- Histogram and density plots of distance to care and income.
- Correlation heatmap for predictor variables.
- Risk score distribution by region.
- Bar charts comparing hospitalization risk by demographic groups.
- Feature importance plots from Random Forest/XGBoost.
- ROC curves for classification models.
- Actual vs. predicted plots for regression models.

**9 Expected Key Insights:**

Based on the analysis, the following key insights are expected:

- Identify regions with high estimated hospitalization risk based on social determinants.
- Provide policy recommendations to improve healthcare access in underserved areas.
- Demonstrate the feasibility of using public SDOH data to estimate hospitalization risk.

## 10 Potential Impediments:

No project progresses from proposal to completion without challenges. As such, we identify the following potential impediments to this project and related plan changes to address these impediments.

First, it is possible that the features we obtained are not sufficient to develop a performant model. If this is the case, we would need to search for additional data to either supplement or replace our current dataset.

Second, given that we have many missing values to impute, it is possible we inadvertently impute values that are wrong enough to skew the models in an incorrect direction. To monitor for this, we will carefully review the model for overfitting and other poor performance outcomes, and if need be, we can use an alternate method for value imputation and repeat the model training, testing, and evaluation.

Finally, it is possible that the patterns observed in one region may not apply directly to other areas due to variations in healthcare infrastructure, policy environments, population density, and cultural factors. Models trained on data from AHRQ region may require recalibration before being applied elsewhere.

## 11 Timeline:

Below is a timeline that identifies deliverables, due dates and conservative estimates of the number of hours per researcher needed to complete this project on time.

| Deliverable Date | Description of Deliverable | Estimated Hours/Researcher |
|---|---|---|
| 3/31 | Proposal Delivered | 18 |
| 4/7 | Data Exploration Report Delivered | 26 |
| 4/14 | Preprocessing & Feature Report Delivered | 26 |
| 4/21 | Model Progress Report Delivered | 26 |
| 4/28 | Model Evaluation Report Delivered | 26 |
| 5/6 | Project Presentation | 42 |
| 5/11 | Written Project Report Delivered | 10 |

**12 Conclusion:**

This preliminary proposal outlines a comprehensive approach to analyzing healthcare accessibility and predicting hospitalization risk using socioeconomic and geographic factors. By leveraging the AHRQ Social Determinants of Health Database and applying robust data science methodologies, we aim to generate valuable insights that can inform healthcare policy and resource allocation decisions. The project aligns with current efforts to address healthcare disparities and improve access to care across diverse communities.

**References:**


1. Artiga, S., & Orgera, K. (2020). Disparities in health and health care: Five key questions and answers. *KFF (Kaiser Family Foundation)*. This source provides broad information on health disparities. https://www.kff.org/racial-equity-and-health-policy/issue-brief/disparities-in-health-and-health-care-5-key-question-and-answers/
2. Am J Manag Care. 2021;27(3):e89-e96. https://doi.org/10.37765/ajmc.2021.88603
3. Agency for Healthcare Research and Quality (AHRQ). (n.d.). Social Determinants of Health (SDOH) Data & Analytics. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
4. Agency for Healthcare Research and Quality – Data Source Documentation https://www.ahrq.gov/sites/default/files/wysiwyg/sdoh/SDOH-Data-Sources-Documentation-v1-Final.pdf
5. National Academies of Sciences, Engineering, and Medicine. (2021). Implementing high-quality primary care: Rebuilding the foundation of health care. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK578537/
6. Bhuiyan, A. R., Fennie, K. P., & Ahmed, S. M. (2020). Exploring the relationship between social determinants of health and healthcare utilization among adults with multimorbidity: a cross-sectional study. *BMC Health Services Research*, *20*(1), 593. https://pmc.ncbi.nlm.nih.gov/articles/PMC7314918/
7. National Academies of Sciences, Engineering, and Medicine. (2012). Living well with chronic conditions: A self-management approach. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK218212/

8. Decision Tree, Random Forest, and XGBoost: An exploration into the heart of Machine Learning. Medium. https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948
9. Cross-validation in machine learning: How to do it right. Neptune AI Blog. Retrieved from https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right