Assessing the Impact of Hospital Ratings and Conditions on Risk-Adjusted Mortality: Evidence from California Hospitals in 2023

Hilda Joseph [1] **Caroline Ngyuen** [1] **Eddie Zhang** [1]

1. Data Description

1.1. Data

The dataset is from data.gov, and refers to a California set about hospital mortality rates for certain conditions and their quality ratings. The data was publicly accessible, and the idea of the paper was established due to the authors' interest in the healthcare field.

Dataset Eddie Zhang wrote the data section

https://catalog.data.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings-8e21f

1.2. Key Variables

Hospital, rating, condition, and risk-adjusted mortality rate The hospitals refer to various hospitals in California. The ratings of the hospitals are either 'worse,' 'better,' or 'as expected' than normal. The condition is the disease the person in the case has. And the risk adjusted mortality rate is the hospital's observed death rate to its expected death rate after accounting for patient-specific risk factors like age. Eddie did the variables.

1.3. Research Question

For each Californian hospital in 2023, how are the hospital ratings and medical conditions associated with the risk adjusted mortality rate and the proportion of deaths out of total cases? Eddie did the research question.

1.4. Scope

The data focuses on Californian hospitals in 2023 because it represents the most recent year available in the dataset. Using the latest year ensures that our analysis reflects the most up-to-date information on hospital performance, operations, and outcomes, making the findings more relevant to our research questions. The dataset includes medical conditions such as: AAA Repair Unruptured, AMI Acute Stroke, Acute Stroke Hemorrhagic, Acute Stroke Ischemic, Acute Stroke Subarachnoid, Carotid Endarterectomy, Esophageal Resection, GI Hemorrhage, Heart Failure, Hip Fracture, PCI Pancreatic Cancer, Pancreatic Other, Pancreatic Resection, Pneumonia. Hilda and Caroline did the scope.

1.5. Challenges in Reading, Cleaning, and Preparing the Data

Since we are working with real-world administrative or performance data, it comes with multiple challenges. Below are key issues and ideas for how to address them:

Small Sample Sizes and Statistical Stability

Some hospitals may have very low case volume for certain conditions or procedures in some years. It is possible that mortality rates derived from small denominators can be highly volatile. Steps need to be taken to normalize the smallest sample sizes. For different case volumes, we will weight each hospital's contribution by the number of cases to reduce the distortion from small-sample hospitals. We will exclude conditions with too few cases that are not statistically meaningful.

Caroline and Hilda wrote challenges.

Numeric Formatting

Some entries might include non-numeric characters, which must be cleaned or parsed carefully. The risk-adjusted mortality rates come in formats such as X.X per 100 cases, we need to clean or adjust the scale to make it consistent We will convert mortality rates, case counts, and volume numbers to numeric types and watch out for cells with text.

Outliers, Extreme Values, and Anomalies

Before modeling, we need to explore distributions and flag extreme outliers is crucial. We will flag suspiciously high or low mortality rates and investigate whether these are valid small-sample effects or reporting errors.

Missing, Suppressed Values - Lack of Cases

Many hospitals have no values for many conditions, meaning there will need to be a lot of adding of NAs to better sort the data. We will identify which fields are empty and exclude them from rate calculations.

Eddie did this specific challenge.

1.6. Steps for Workflow

Read In with Care

We will use a library depending on the language that tolerates missing or malformed entries and reads numeri-

Hilda did this.

cal/string variables.

Initial Sanity Checks

We plan to count missingness per column, check ranges of numeric variables, look for impossible values (i.e. negatives, greater than 100 percent mortality), narrow down the dataset to the year 2023 only, and inspect sample sizes.

Caroline did this.

Standardized Column Names and Formats

It is important to unify naming (e.g. "AMI" vs "Acute Myocardial Infarction") and convert rates to consistent numeric scale, coerce types.

Hilda did this.

Model-Ready Dataset

Ensure to drop or flag residual missing entries, impute if justifiable, and divide into training/test splits if doing an explanatory modeling.

Documentation

Caroline did this.

We will keep track of cleaning rules for the dataset (e.g. how missing values are treated) to ensure transparency and reproducibility.