Hospital Cost & Reimbursement Analysis (2018–2022)

1. 1. Introduction

Hospitals often list standard charges for medical procedures that differ significantly from the actual reimbursements they receive from payers (Medicare, Medicaid, private insurers). This discrepancy impacts hospital financial health, payer negotiations, and patient out-of-pocket costs. Understanding these gaps helps stakeholders make informed decisions and promotes pricing transparency.

1. 2. Objective

Analyze the charge-to-payment ratios for Medicare-certified hospitals across the United States from 2018 through 2022. Key objectives include identifying temporal trends, distribution characteristics, and enriching the dataset with demographic metrics.

1. 3. Research Questions

• How has the average charge-to-payment ratio evolved from 2018 to 2022?

• What is the distribution of these ratios each year, and are there significant outliers?

• Which hospitals and states have the highest and lowest average ratios?

• How does hospital demographic context relate to charge-to-payment ratios?

1. 4. Data Sources

4.1 CMS Hospital Cost Reports (2018–2022)  
• Source: CMS Provider Summary Data (HOSPITALS) portal.  
• Description: Annual cost report CSV files with total billed charges, total Medicare payments, and facility metadata.  
• Key Fields: CCN, Total\_Charges, Total\_Payment, Report\_Fiscal\_Year, State.

4.2 CMS Hospital General Information  
• Source: CMS Hospital General Information dataset.  
• Description: Facility-level metadata including hospital name, type, and ownership.  
• Key Fields: CCN, Hospital Name, Hospital Type, Hospital Ownership.

4.3 County-Level Demographic Data (SAHIE)  
• Source: U.S. Census Bureau – Small Area Health Insurance Estimates (SAHIE).  
• Description: Annual county-level health insurance coverage rates.  
• Key Fields: County FIPS, Percent Insured, Age Groups, Income Brackets.

1. 5. Data Cleaning & Preprocessing

• Dropped irrelevant columns and filtered data to 2018–2022.

• Standardized column names and converted monetary values to numeric types.

• Handled missing or erroneous entries by removing invalid rows.

• Calculated charge\_to\_payment\_ratio = Total\_Charges / Total\_Payment.

• Filtered out extreme ratio values (>=100) and invalid entries (NaN, Inf).

1. 6. Methodology

• Temporal Analysis: Computed average ratio per year and plotted a time series.

• Distribution Analysis: Generated boxplots and histograms for yearly ratio distributions.

• Outlier Detection: Identified hospitals in the top/bottom 1% of ratios.

• Demographic Enrichment: Merged SAHIE data to add county-level insurance metrics.

• Dataset Preparation: Saved enriched dataset for further analysis.

1. 7. Data Analysis & Results

7.1 Temporal Trends:  
– The average charge-to-payment ratio increased from approximately 2.1× in 2018 to 2.3× in 2022, with variability peaking during 2020–2021.

7.2 Ratio Distribution:  
– The majority of hospitals fall between 1× and 10×; distributions are right-skewed with long tails.

7.3 Demographic Enrichment:  
– Enriched dataset includes county-level percent insured.  
– Prepared for future correlation analysis between insurance coverage and markup ratios.

1. 8. Notebook 04: County Demographics Enrichment

Completed:  
• Imported SAHIE demographic data and merged with the hospital dataset using county FIPS codes.  
• Added metrics such as percent insured at the county level.  
• Saved the enriched dataset to CSV for further community-level analysis.

1. 9. Conclusion

This analysis highlights consistent markup behaviors across U.S. hospitals, with significant variability during the COVID-19 period and right-skewed distributions. Demographic enrichment sets the stage for exploring the relationship between insurance coverage and hospital markup strategies.

1. 10. Limitations & Future Work

• Data limited to Medicare-certified hospitals; private payer and uninsured patient data are excluded.

• Facility metadata assumed static; multi-year merges could improve accuracy.

• Extreme ratio filtering may exclude valid high-markup cases; refine outlier handling in future.

• Demographic correlation and RUCC classification are pending for deeper geospatial insights.