model-visualization

1 Project Motivation

1.1 Background on Dialysis and Proposition 29

Dialysis is a lifesaving treatment that removes waste from blood, acting as an artificial kidney for those with chronic kidney disease. In recent years, California has seen multiple attempts to increase regulations on dialysis facilities through ballot initiatives:

- Proposition 8 in 2018
- Proposition 23 in 2020
- Proposition 29 in 2022

Proposition 29, which failed to pass, aimed to establish increased regulations for both staffing and operations for the roughly 600 dialysis facilities in California, an estimated $3.5 billion industry. A key provision of the proposition required the presence of a physician or licensed practitioner during all treatment hours, potentially increasing each facility's costs by several hundred thousand dollars annually.

1.2 Debate Surrounding Dialysis Regulations

- **Proponents** argue that increased regulations improve patient safety and quality of care.
- **Opponents** contend that the increase in healthcare costs is unwarranted and would limit care coverage by overwhelming facilities with costs and potentially forcing them to close.

1.3 Project Scope and Significance

Our project explores trends in:

1. Dialysis facility access
2. Quality of care
3. Ballot results in California

Using publicly available data from:

- Center for Medicare and Medicaid Services
- Aggregated election results from California's Secretary of State

We analyze associations between dialysis care and voting behaviors, specifically those related to recent statewide ballot initiatives designed to regulate California's multibillion-dollar dialysis industry.

1.4 Relevance and Novelty

- This project uses a novel approach to study an area of public interest relevant not just to California but the entire country.
- To our knowledge, this is the first project of its kind to explore the possible association between voting patterns and the quality of dialysis care received.
- The majority of dialysis treatments in California are covered by Medicare, making the Medicare and Medicaid datasets particularly relevant to our analysis.

1.5 Broader Context

Investigative reporters, patient advocacy groups, and labor organizations have spent significant resources over the past decade to raise public awareness of the dialysis industry and its need for regulation. Our project contributes to this ongoing discussion by providing data-driven insights into the relationship between dialysis care quality and voting behavior.

2 Project Research Questions

Our research questions are divided into two categories: primary and secondary. The primary question serves as the main focus of our analysis, while secondary questions provide additional insights through further investigation of the data.

2.1 Primary Research Question

Is the quality of care of dialysis facilities correlated with voting in favor of or against dialysis industry regulation?

2.1.1 Key Assumptions

1. The relationship between Quality of Care and Voting Behavior is not confounded.
2. A vote in favor of any of the three propositions (Prop 8 in 2018, Prop 23 in 2020, Prop 29 in 2022) can be interpreted as support for dialysis industry regulation.

2.1.2 Quality of Care Metrics

To test this relationship under the outlined assumptions, we approximate the quality of care using the following metrics:

- Five-star rating
- Patient experience rating
- Facility mortality rate
- Number of available dialysis stations
- Staff rating
- Hospital readmission categorization (Worse than Expected, As Expected, Better than Expected)
- Profit/non-profit designation
- Parent company affiliation/independence

2.1.3 Facility Categorization

Observations of Quality of Care metrics in our data are categorized by:

- Year
- County
- City
- Profit/Non-profit designation
- Parent company affiliation/independence

2.2 Secondary Research Questions

1. What is the geographic coverage of dialysis facilities in California?
2. Is there any correlation between organizational structure (chain-owned, profit vs. non-profit) and the quality of care?
3. Is there any association between the parent company of dialysis facilities and the quality of care?

3 Data Sources

3.1 Primary Data Sources

3.1.1 Center for Medicare and Medicaid Services (CMS) Quarterly Dialysis Facility Compare Dataset

- **Source Details:**
  - Type: .zip & .xlsx
  - Years: 2017-2024
  - Combined Size: 5,684 rows, 173 columns
  - Link: CMS Dialysis Facilities Data
- **Key features**:
  - Star ratings for dialysis facilities
  - Patient experience metrics
  - Quality of care metrics
- **Insights provided**:
  - Patient satisfaction
  - Clinical outcomes
  - Doctor-patient communication
  - Hospitalization rates
  - Treatment effectiveness
- **Rating calculation**:
  - Patient experience star rating derived from bi-annual patient surveys
  - Facility ratings based on metrics including:
    - Unplanned hospital readmissions
    - Total and expected transfusions
    - Ratio of deaths to expected deaths
    - Waste removal efficiency across patient types

3.1.2 California Secretary of State's (SOS) Statement of Vote for Ballot Measures

- **Source Details:**
  - Type: .xlsx
  - Years: 2018, 2020, 2022
  - Combined Size: 7,122 rows, 7 columns

  - Link: CA SOS Statewide Election Results

- **Elections covered**: November 2022, 2020, and 2018
- **Focus**: Propositions regarding dialysis clinic requirements
- **Geographic levels**:
  - Counties
  - Sub-counties:
    * Congressional districts
    * State senate districts
    * State assembly districts
    * Cities

3.2 Secondary Data Source

3.2.1 California Health and Human Services Specialty Care Clinic Complete Data Set

- **Source Details:**
  - Type: .xlsx
  - Years: 2013-2023
  - Combined Size: 6,605 rows, 143 columns
  - Link: CHHS Specialty Care Clinic Data
- **Time range**: 2013 through 2023
- **Purpose**: Supplement CMS dataset with additional geographic data
- **Additional facility-level features**:
  - Senate district
  - Congressional district
  - Latitude and longitude

3.3 Data Integration and Analysis Potential

- Multiple geographic levels allow for aggregation and analysis at various scales
- Combination of clinical data (CMS) with voting data (SOS) enables exploration of potential correlations between care quality and voting behavior
- Supplementary geographic data enhances spatial analysis capabilities

4 Data Manipulation Methods

Our workflow was broken down into five stages:

1. Data Collection
2. Data Preparation
3. Database Management
4. Exploratory Data Analysis
5. Statistical Analysis

4.1 Data Collection and Preparation

4.1.1 CMS Dialysis Facility Dataset

4.1.1.1 Organization and Import

- Dataset structure: .zip files (one per year), containing multiple Excel files
- Focus: Excel files relevant to facility general information, ratings, and patient survey results
- Import result: Two separate parquet files at the facility level

    1. Patient survey responses
    2. Facility ratings and measurements

4.1.1.2 Challenges and Solutions

1. Inconsistent File Naming Conventions

    - Issue: 2021 files named differently (e.g., patient survey data file named '59mq-zhts')
    - Solution: Created a list of exact file names for selection, rather than using pattern matching

2. Missing Data

    - Expected missing data: Survey non-responses
    - Unexpected missing data: Administrative errors (e.g., missing columns in recent ICHPS raw data files)
    - Solution for specific cases: Simple imputation during analysis (e.g., substituting 2018 'nan' values with 2019 values at the facility level)

4.1.2 SOS Ballot Data

4.1.2.1 Import and Selection

- Data imported via URL for each relevant proposition year (2018, 2020, 2022)
- Selected columns containing ‘Kidney’ or ‘Dialysis’ for analysis
- Geographic column manipulation:
  - Renamed columns
  - Backfilled rows to address multi-level index (sub-counties under counties)
- Final output: Single ballot data parquet file
  - Includes year column
  - Count and sub-county vote counts for each Dialysis Requirements Initiative proposition

4.1.2.2 Challenges and Solutions

- Inconsistent naming conventions across years
  - 2020 and 2022: ‘County Supervisorial’
  - 2018: ‘Supervisorial District’
- Solution: Standardized naming across all years

4.1.3 CHHS Specialty Care Clinic Complete Data Set

4.1.3.1 Import and Alignment

- Downloaded Excel files for 2013 through 2023 (one per year)
- Main challenge: Aligning pre-2018 data with 2018-forward structure
- Process:
  1. Separated data into two dataframes: 2013-2017 and 2018-2023
  2. Used CHHS mapping dictionary to rename 2013-2017 columns
  3. Ensured consistent data types across both dataframes
  4. Merged dataframes using outer join on common columns
  5. Dropped rows with missing FAC_NO (facility data)

4.2 Database Management

4.2.1 Data Merging and Standardization

- Standardized data types and column names across all datasets
- Merged datasets:
    1. CMS facility rating dataset with CMS patient survey dataset
    2. Filtered CHHS dataset (dialysis clinics only) with merged CMS data
    3. Reshaped CMS and CHHS data by geographic level
    4. Merged geographic-level data with SOS Ballot Measures dataset

4.2.2 Final Output

- Two parquet files:
    1. Data aggregated at city level
    2. Data aggregated at assembly district level

4.2.3 Custom Relational Database System

- We developed a custom Python-based relational database system to centralize our datasets and facilitate efficient data access and analysis. Key features include:
    1. **Table and View Structure**: Distinct tables for datasets with multiple views for focused data access.
    2. **Dynamic View Creation and Merging**: On-the-fly creation of custom views and combination of multiple views for complex analysis.
    3. **Conditional Querying**: User-defined conditions for precise data retrieval and filtering.
    4. **Efficient Data Access**: Quick and reliable access across the entire database.
    5. **Code Quality**: Object-oriented design with consistent naming conventions for improved maintainability and adaptability.
- For detailed functionality, refer to the included database demo (Milestone_1/004_data-processing-scripts/002_clean-raw-data/database_demo.ipynb).

5 Analysis and Insights

This project employed a Bayesian approach to investigate the relationship between dialysis facility quality metrics and voting patterns on dialysis-related propositions in California.

By integrating these datasets, we were able to explore possible relationships between dialysis facility quality metrics and voting outcomes on dialysis-related propositions, an analysis that would not have been possible with either dataset alone.

To maximize the granularity of our analysis, we chose to focus on city-level data, which provided more detailed vote counts compared to assembly district level data. We encoded voting outcomes as the percentage of "Yes" votes in favor of the propositions, allowing for a nuanced examination of support for dialysis industry regulation across different localities.

This novel approach enabled us to investigate how various factors related to dialysis care quality might influence public opinion and voting behavior on healthcare policy initiatives, potentially offering valuable insights for policymakers, healthcare providers, and voters alike.

5.1 Analysis Steps

5.2 Data Preparation

Before modeling and analysis, we underwent several data preparation steps:

1. Data Loading: We loaded the merged dataset containing both CMS facility data and ballot measure voting outcomes using the `read_parquet` function.
2. Data Imputation: For the year 2018, we imputed missing values for two variables (`ich_cahps_survey_of_patien` and `patient_hospital_readmission_category`) using 2019 data. This was done to ensure consistency across years and minimize data loss.
3. Data Transformation: We performed several transformations on the dataset:
    - Converted relevant variables to numeric format (e.g., `mortality_rate_facility`, `n_dialysis_stations`, `staff_rating`, `five_star`, `patient_experience_rating`).
    - Created an ordered factor for `hospital_readmission`.
    - Calculated the vote percentage in favor of regulation for each facility.
4. Data Filtering and Aggregation: We filtered the dataset to include only years 2018, 2020, and 2022. We then aggregated the data at the facility level, summarizing vote counts and percentages, and selecting relevant facility characteristics.
5. Data Cleaning: We removed any rows with missing values to ensure a complete dataset for modeling.

This processed dataset, `filtered_city_df`, includes variables such as year, provider number, profit status, chain organization, county, city, facility ratings, mortality rate, staff rating, patient experience rating, number of dialysis stations, hospital readmission category, and vote percentage. This comprehensive dataset forms the basis for our subsequent modeling and analysis of factors influencing voting outcomes on dialysis clinic regulation.

5.2.1 Model Construction

We constructed a Bayesian multilevel model using the brms package in R. This model allowed us to account for the hierarchical nature of our data (facilities nested within cities and counties) while examining the relationship between facility quality metrics and voting outcomes.

5.2.2 Posterior Predictive Checks:

We performed posterior predictive checks to assess model fit and explore relationships between key variables and voting outcomes.

5.2.3 Visualization of Effects:

We created visualizations to illustrate the effects of key variables on voting outcomes, such as staff rating and mortality rate.

5.3 Insights

5.3.1 Staff Rating Impact:

Our analysis revealed a negative relationship between staff ratings and the percentage of votes in favor of dialysis regulation. This suggests that areas with lower-rated dialysis facility staff were more likely to support increased regulation. The effect of staff rating varied across counties, with some counties showing stronger negative relationships than others.

5.3.2 Mortality Rate Influence:

We found a positive relationship between facility mortality rates and support for regulation. As mortality rates increased, the predicted vote percentage in favor of regulation also increased. This suggests that voters in areas with higher mortality rates at dialysis facilities were more likely to support increased oversight.

5.3.3 Patient Experience Rating:

Interestingly, our analysis showed a positive relationship between patient experience ratings and support for regulation. This counterintuitive finding suggests that even in areas where patients report better experiences, there is still support for increased regulation.

5.3.4 Five Star Rating and Stations per Facility:

The associations of these metrics on voting behavior were weaker compared to staff ratings, mortality rates, and patient experience ratings. The estimated effects suggested by the posterior predictive checks were clustered around zero.

5.3.5 Chain Organization Effects:

The posterior predictive check for chain organizations showed varying levels of support for regulation across different dialysis chains, indicating that organizational factors may play a role in shaping public opinion or voting behavior.

5.3.6 Facility Size Considerations:

The number of dialysis stations (a proxy for facility size) showed a slight positive relationship with voting in favor of regulation, suggesting that areas with larger facilities might be more supportive of increased oversight.

5.4 Challenges and Limitations

5.4.1 Data Granularity:

While we had facility-level data for quality metrics, aggregating voting data at the city and county level creates a mismatch in granularity, potentially obscuring some finer-grained relationships.

5.4.2 Temporal Alignment:

Our analysis assumed that facility metrics from a given year directly influenced voting in that year's proposition. However, there may be lag effects or longer-term trends that our current model doesn't capture.

5.4.3 Confounding Factors:

Despite our efforts to control for various factors, there may be unmeasured confounders influencing both facility quality and voting patterns that our model doesn't account for.

5.4.4 Causal Interpretation:

While our model reveals associations between facility metrics and voting patterns, caution should be exercised in interpreting these relationships as causal. Further research, possibly including quasi-experimental designs, would be needed to establish causal links.

In conclusion, our analysis provides novel insights into the complex relationship between dialysis facility quality and public support for industry regulation. The varying effects across counties and organizations highlight the importance of considering local contexts when interpreting these results. While we've uncovered several interesting patterns, the complexity of the issue and limitations in our data suggest that further research is needed to fully understand these relationships and their policy implications.