Assessing the Performance of Dialysis Facilities Participating in the End Stage Renal Disease (ESRD) Network Program

Background

Chronic Kidney Disease (CKD) affects about 30 million US adults, and kidney diseases are the ninth leading cause of death in the US. About 96% of people with kidney damage or reduced kidney function are unaware of having CKD. Every 24 hours, more than 300 people begin dialysis treatment for kidney failure; and 75% of new cases of kidney failure in the US are caused by diabetes and high blood pressure. In 2015, total Medicare costs for CKD and ESRD treatment in the US were $98 billion.

Dialysis is a process of filtering the blood, and is needed for patients with End Stage Renal Disease (ESRD) due to filed kidney function. Hemodialysis and peritoneal dialysis helps control blood pressure and maintain mineral balance in the blood. The process involves access to the patient's bloodstream using a 'vascular access' site, usually a fistula, graft, or a catheter, and patients undergoing hemodialysis are at high risk of infections such as Staphylococcus aureus, bloodstream infections, and viral hepatitis. In the US, there are approximately 7,000 Medicare- approved facilities participating in ESRD Networks, and CMS monitors and evaluates the performance of dialysis facilities through quality improvement initiatives.

Motivation and objectives

The objective of this project is to examine publicly available Medicare data for US dialysis facilities participating in the End Stage Renal Disease (ESRD) Network Program. The data consists of numerous sources and measures related to quality improvement, network performance, and patient morbidity and mortality. The motivation to use this data was to gain hands-on experience with applying the technologies associated with the R programming language to access and extract meaningful information and insights from different data sources pertaining to Dialysis facilities. Data sources

available to public health practitioners often vary in quality and formats, and a primary motivation to embark on this this project was to gain skills applicable in the workplace, as well as an understanding of emerging data-science related learning needs of the public health workforce.

The variables used in this project were extracted from the Dialysis Facility Compare datasets[2], and the data available is related to facility network information (facility address, network, chain owned etc.), and several performance-score related measures (five star rating, total performance score, number of dialysis stations, facility mortality rate, hospitalization/ readmission rates, dialysis adequacy score, blood stream infection score, facility transfusion score, anemia management score, vascular access score). Additionally, data for patient demographics and transplant patterns for regression models were extracted from the Dialysis Facility Report[3] dataset. Chronic kidney prevalence data was obtained from

The initial research questions related to this project were focused on descriptive analysis of the variation in quality of clinical care, patient characteristics, morbidity and mortality, and transplantation patterns in dialysis facilities. There was also a focus on geocoding facility addresses to perform analysis in geographic variation. The main progression in the research questions was that they moved from a purely descriptive standpoint to a predictive model, while retaining the original intent of conducting an exploratory analysis. Using the application of various technologies such as regression modeling and cluster analysis, the focus of the project turned to create models, and identify clusters in the data. Below are the research questions guiding the project:

1. What is the distribution of patient mortality and morbidity within performance score categories and five-star rating categories?
2. What are the predictors of total performance scores in dialysis facilities?

3. Are there any specific cluster patterns in these data? An assumption would be that observed clusters would align with observed categories of total performance scores/ five-star rating
4. What are some factors associated with the differences in the number of transplants in networks?
5. What is the sentiment of public discussion about Dialysis on social media sites, news articles, and patient forums?

Methods

The data analysis was conducted by using base R programming language, and R packages. Data sources are listed below:

1. **Dialysis Facility Compare (DFC)** data is available at https://data.medicare.gov/data/dialysis-facility-compare
    - DFC datasets listed below were accessed using SOCRATA API available on the Medicare website. API access requires creating an account and obtaining apptokens, and using Rsocrata package
        - Dialysis Facility Compare - Listing by Facility
        - ESRD QIP - Complete QIP Data - Payment Year 2018
        - Patient survey (ICH CAHPS)
2. **Dialysis Facility Report** data is available at: https://www.dialysisdata.org/content/dialysis-facility-report-data
    - This dataset was primarily used to extract variables related to Race, Ethnicity, ESRD cause, number of kidney transplants
    - Read in using Readr package
3. **US Chronic Disease Indicators: Chronic Kidney Disease Prevalence data** is available at: https://catalog.data.gov/dataset?tags=chronic-kidney-disease
    - This dataset was used to extract the 2016 crude prevalence rates of Chronic Kidney Disease from BRFSS data
    - Data was imported into DB Browser for SQLite, and read in using RSQLite package:
4. **Zip Code XY coordinates** obtained from R "zipcode" package
5. Twitter
6. Patient forum data accessed from: https://patient.info/forums/discuss/browse/kidney-failure-and-ckd-1300

Other R packages used:

1. DPLYR: Used for data manipulation
2. GGPLOT2, MAPS, Leaflet: Data visualization
3. RTWEET: Access Twitter data using API

4. TM, Wordcloud, SentimentAnalysis, Tidytext: Text analysis
5. Zipcode: Geocoding
6. RSocrata (API for DFC data)
7. Rvest: Webscraping

Procedures

After importing the raw datasets, variables for the final dataset were extracted from each data source, and rate and measure variables were converted to numeric type. Dialysis facility data sets were merged primarily on the 'provider number' identifier. Zip codes were joined to the facility address zip codes using the "zip" identifier, and CKD prevalence data was joined on the "state" identifier. Continuous variables for performance score, patient facility rating, and number of dialysis stations were categorized for data summarization and visualization. Exploratory analysis was conducted to summarize and visualize the distribution of variables of interest, and to generate histograms, bar plots, and maps. Additionally, cluster analysis was conducted to identify clustering patterns in the data, and regression analysis was conducted to create a model to predict total performance scores in facilities. Finally, sentiment analysis of data collected from social media, news articles, and patient forum discussions was conducted to identify discussion sentiments.

Summary of findings

Exploratory analysis results show that most dialysis facilities (88%) are chain-owned, and networks contain a mix of facilities with poor to good performance scores. A majority of the patient population is African American (mean percentage 60%), and White (mean percentage 33%). Most patients have Diabetes as the primary reported cause of ESRD (44%), and about 28% report hypertension as the primary cause of ESRD. Total performance score is most strongly predicted by facility readmission rate, standardized hospitalization ratio, transfusion score, blood stream infection measure score, dialysis adequacy score, and combined vascular access score (model R-squared= 0.85). The current five-star rating method seems to be a robust classification of facility characteristics, and cluster analysis of the

data do not identify any other specific clustering patterns. The number of kidney transplants is strongly correlated with the number of dialysis stations in facilities, suggesting that an adequate dose of dialysis is essential to sustain patient health for receiving transplants. The public discourse related to dialysis on social media and news show positive sentiment, but text analysis of discussions of patient forums project negative sentiment.

References

1. Dialysis Facility Report data: https://www.dialysisdata.org/content/dialysis-facility-report-data
2. Dialysis Facility Compare Datasets: https://data.medicare.gov/data/dialysis-facility-compare (SOCRATA API)
3. https://www.niddk.nih.gov/health-information/kidney-disease/kidney-failure/hemodialysis
4. https://www.cdc.gov/dialysis/patient/index.html
5. US Chronic disease indicators:https://catalog.data.gov/dataset?tags=chronic-kidney-disease - JSON export available
6. http://esrdnetworks.org/membership/esrd-networks
7. https://www.cms.gov/Medicare/End-Stage-Renal-Disease/ESRDNetworkOrganizations/