Zaid Nasr Capstone Proposal DATS4001

Objective

Use ICD9/ICD10 codes obtained via the ICD-API v2 to identify (1) the distribution of Type 1 vs. Type 2 diabetes, (2) the prevalence of key diabetic complications (kidney disease, neuropathy), and (3) rates of diabetes-related hospital readmissions. Then, map these metrics across different regions and demographic groups to highlight treatment inequities.

Impact

By unveiling where and how diabetes management gaps are most prevalent, this project will enable healthcare providers and policymakers to allocate resources more effectively. Mapping inequities in prevalence, complications, and readmissions will highlight regions and demographic groups in greatest need of interventions.

Data Collection Methodology

This project will utilize the ICD-API v2 provided by the WHO to access ICD9 and ICD10-coded hospital claims data. By programmatically retrieving data using OAuth2 authentication, I will query ICD-coded records related to diabetes, extracting diagnosis codes, patient demographics, hospitalization details, and readmission rates. This approach ensures real-time access to the most updated and structured ICD-encoded medical records.

Datasets

ICD9-Coded Claims/Records

- Source: Retrieved via the ICD-API v2
- Data Types: Diagnosis codes, basic demographic information, and hospitalization details coded under the ICD9 system.
- Reliability: Accurate coding depends on clinical documentation and hospital billing practices, which can lead to undercoding or upcoding. Rigorous data cleaning and cross-checking of records can help maintain data quality.

ICD10-Coded Claims/Records

- Source: Retrieved via the ICD-API v2
- Data Types: Diagnosis codes, patient demographics, and treatment episodes encoded according to ICD10 standards.
- Reliability: ICD10 offers more specificity than ICD9, though clinical documentation errors can still occur. Careful validation and a solid data-cleaning protocol will improve my project's overall reliability.

Approach

I will begin by gathering all ICD9 and ICD10 claims that mention diabetes using the ICD-API v2. First, I'll label each record by whether it refers to Type 1 or Type 2 diabetes, any complications, or any hospital readmissions. Then, I'll clean the data by removing duplicates, addressing any missing details, and converting ICD9 codes to their ICD10 equivalents so everything aligns correctly.

Next, I'll explore the cleaned dataset to see how common each type of diabetes is, how often certain complications occur, and how frequently patients end up back in the hospital. If the dataset includes location information, I'll also look at the data by region to spot any patterns or areas with bigger problems. I'll use Python-based statistical analysis and visualization tools (such as Pandas, NumPy, and Matplotlib) to generate insights.

Finally, I'll compile these findings into a clear report, complete with charts (and maps if available) to highlight diabetes treatment gaps and inequities. I'll also document my steps in detail so others can follow or expand on this work later.

Revised Timeline

- (2 Weeks) Data Gathering – Retrieve ICD9 and ICD10 claims data via ICD-API v2, ensuring proper API authentication and access.
- (2 Weeks) Data Cleaning & Code Mapping – Remove duplicates, handle missing details, and convert ICD9 codes to ICD10 equivalents for consistency.
- (3 Weeks) Exploratory Analysis – Analyze diabetes prevalence, complications, and hospital readmission patterns; create visualizations.
- (2 Weeks) Spatial or Group Analysis – Use geographic or demographic identifiers (if available) to identify healthcare disparities.
- (1 Week) Reporting – Compile findings into a concise report with relevant charts/maps.
- (1 Week) Final Presentation – Deliver a presentation summarizing the project.

Possible Issues & Mitigation Strategies

- Data Privacy (HIPAA Compliance) – I will only use anonymized data and handle it securely to ensure compliance.
- ICD9 to ICD10 Conversion Challenges – I will use established code-mapping guides to ensure accurate translation between code systems.
- Geographic Linking Challenges – If I use geospatial mapping, I will apply careful data validation techniques to prevent misclassification.

Predictive Model Component

In addition to analyzing disparities in diabetes care, I plan to develop a predictive model to estimate how a community's healthcare coverage may change over time. This model will assess whether a region is likely to experience an improvement or decline in healthcare access and quality based on historical trends and socioeconomic indicators.

Predictive Model Approach

1. Data Inputs:
    - Historical diabetes-related healthcare coverage scores
    - Changes in income, insurance coverage, hospital infrastructure, and demographic shifts
    - Previous rates of complications and readmissions
2. Machine Learning Methods:
    - Logistic Regression to predict binary outcomes (improvement vs. decline)
    - Random Forest for feature importance analysis
    - Time-Series Forecasting (ARIMA or LSTM) for long-term trends
3. Output:
    - A "coverage score" projection for each region, indicating whether healthcare access for diabetes patients is expected to improve or worsen.
    - Visualizations such as heatmaps and trend lines to illustrate regional predictions.

By incorporating predictive analytics, this project will not only highlight existing disparities but also provide actionable forecasts.