Report

Title: Investigating Causes, Demographic Characteristics, and Prevalence of Diabetes the United States

IE Faculty/Research/Advisor: Dr. Shihao Yang
Student: Qiming Wei

1. Abstract

Diabetes is a growing concern among youth and young adults, particularly in the where the prevalence of both Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM) is rapidly increasing. This study proposes a retrospective longitudinal analysis using Medicaid claims data to explore the causes and effects of diabetes, identify key demographic characteristics, and analyze the trends in diabetes prevalence among Medicaid-enrolled youth and young adults. We will be focusing on a 14-year period (2005-2018) and applying statistical models in order to uncover critical causes of diabetes based on factors such as age, race/ethnicity, and geographic location. The research data resources would be mostly based on sponsor CMS(center for medicaid & Medicare Service).

For this Project at this stage, the challenge arises due to the inconsistency in table structures, size of data, and data formats for each year, inpatient and outpatient records. By implementing SQL stored procedures, loops, and validation techniques, we achieved large progress on extraction, standardization, and unification of these records. This report outlines the methodologies employed and current results obtained.

2. Introduction

2.1 Problem

Healthcare databases, particularly those spanning multiple years, often exhibit heterogeneous structures due to evolving schema standards and differing use cases. For this study, the challenge is to extract patient records with specific diagnostic codes ("target codes") from tables representing inpatient and outpatient records from 2005-2018. Due to the properties of the database, we have to gather data from a minimum of 6 different tables that are for both inpatient records and other service records and time periods from 2005–2012, 2013–2015, and 2016–2018.

2.2 Objective

To develop a scalable and efficient process to:

1. Identify and extract patients with target diagnostic condition/codes and condition:
    a. **Medication Criteria**:
        i. Sulfonylurea, insulin, alpha-glucosidase inhibitors, meglitinide, amylin analogs, or dipeptidyl peptidase inhibitors.
        ii. Biguanide (metformin), thiazolidinedione, GLP1-RA, GLP1-GIP dual agonist, or SGLT2 inhibitors (this must be paired with another criterion to meet the requirement).
    b. **Diagnostic Codes**:
        i. ICD-9-CM: 250.x0, 250.x2, 357.2, 366.41, 362.01–362.07.
        ii. ICD-10-CM: E08., *E09*., E11., *E13*., O24.1*, O24.3*, O24.8*.
2. Two-year condition:
    a. A patient is identified as having incident diabetes when any 2 of the following criteria are met at least once within a 24-month period OR the same criterion is met twice on separate days within the same 24-month period, with the exception of medications as noted.
3. Standardize records across diverse table structures.
4. Validate the accuracy and consistency of the extracted data.
5. Analyze the cause and effect between the diabetes and the patient's demographic conditions.

---

3. Methodology

3.1 Data Challenges

- **Structural Variations**: All tables in the database have various column names, format of columns, numbers of each category of columns (e.g some have 2 columns for target code, some have 9 and some have 12 columns) and structures and data type (e.g some use int some use string, some diagnostic code record using ICD-9 and other uses ICD-10 or something else).
- **Heterogeneous Data Periods**: Data for 2005–2012, 2013–2015, and 2016–2018 are stored separately, with structural discrepancies within these periods.
- **Ambiguity in Records**: Missing, duplicated, or conflicting patient data required additional validation.
- **Time Theory**: The database holds an extremely large amount of data, it is super important to optimize the algorithm to reduce the run time. The queries for large data normally take days and even weeks to run.

3.2 Approach

Extract Target Diagnostic Codes:

Step 1: Table Segmentation and Initialization

Other than spending time studying and knowing each of the tables well, to manage the database effectively, we first need to create intermediate tables and standardize the structures.

- **Create Intermediate Tables for Each Time Period**:
  - Extract raw data into separate tables based on the time periods (2005–2012, 2013–2015, 2016–2018) and patient types (inpatient and outpatient).
  - Ensure each intermediate table includes relevant columns for initial analysis.
- **Standardize Column Names and Order**:
  - Align column names and ensure consistent ordering across all tables to facilitate merging and subsequent processing.
  - Use a predefined schema to normalize structural discrepancies.
- **Filter by Target Diagnostic Codes**:
  - Apply filtering criteria to retain only records matching the specified diagnostic codes.

These steps create 6 clean and consistent intermediate tables, ready for further processing and standardization.

Step 2: Standardizing Table Structures

To address structural variations and ensure consistency across tables:

1. **Create Unified Column Names and Formats**:
   - Align column names and data types across all tables to match a standardized schema.
   - Ensure columns such as patient_id, state_code, diagnostic_code, and service_date are consistently formatted.
2. **Compute Min and Max Values for Validation**:
   - For demographic columns (e.g., dates, gender, birthdate), calculate the minimum and maximum values.
   - These metrics will help identify outliers and ensure data validity in later steps.
3. **Merge Diagnostic Codes into a Single String**:
   - Use a concatenation process to consolidate multiple diagnostic codes for a single patient.
   - This ensures compatibility when merging records

Step 3: Combining Tables

To unify data across inpatient and outpatient records from each year:

- **Create a combine table**:
  - Create a new table with all necessary columns.
  - Insert the corresponding data from the 6 tables from the previous step into the new create table.
- **Validate Merged Data**:

    - Check the row number to ensure data integrity

Step 4: Validating Data Consistency

To ensure data accuracy before applying the two-year condition:

1. **Compare Min and Max Values**:
    - Identify records where the min and max values for a column are inconsistent.
    - Flag these records with missing or conflicting data (Max = ! Min).
2. **Clean Data:**
    - Manually work a case study to analyze the problems
    - Clean the data with problem and move on to next step

Step 5: Applying the Two-Year Condition

As the final step, determine if patients meet the two-year diagnostic criteria:

1. **Identify Records Meeting Criteria**:
    - For each patient, check if:
        - At least two different diagnostic or medication criteria are satisfied within a 24-month period.
        - Alternatively, the same criterion is satisfied twice on separate dates within the 24-month period (this can apply the same coding logic together with the previous one)

    Loops and conditional logic were implemented (as the most quick method) to iterate through data records for identifying patients with diagnostic codes that appear within a two-year window. For each patient:

    - Fetch their records ordered by time.
    - Compare each new record's service date with the previous record's date.
    - If the difference in days is less than or equal to two years, mark the record as satisfying the constraint.
2. **Select the target patient population:**
    - Create a table that only stores all the distinct patient_id which pass all of our conditions.

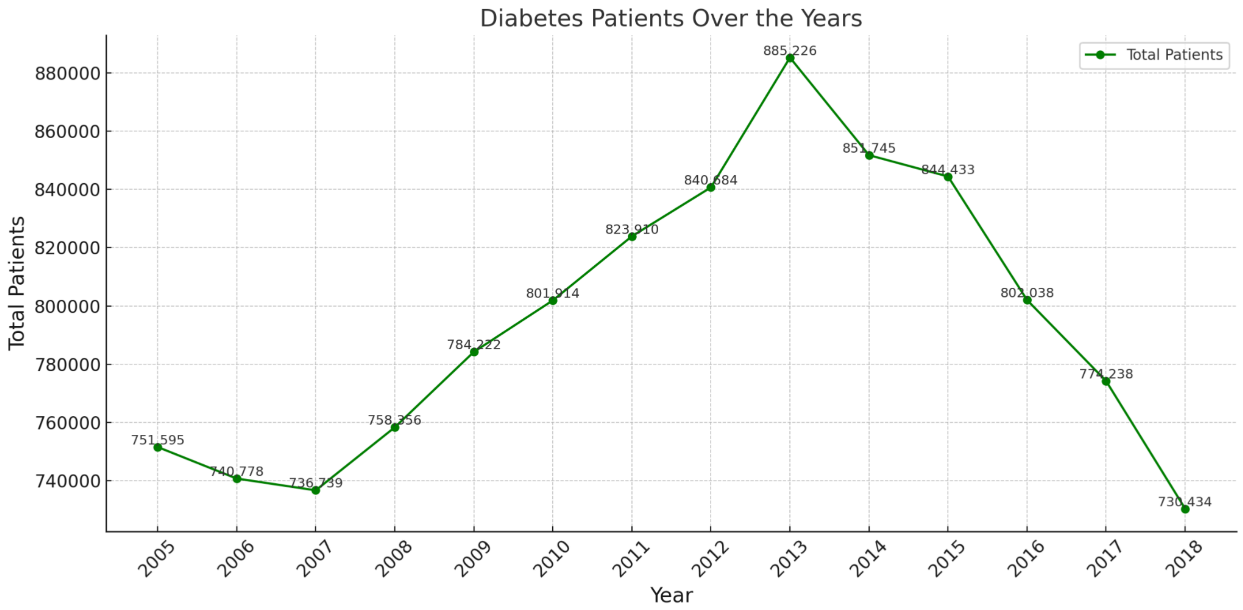
---

4. Current Results

We were able to create the table for all target patient_id in database, here is a example for Total Patients in GA with Diabetes Over the Years (2005–2018):

Year	Total Patients
2005	751,595
2006	740,778
2007	736,739
2008	758,356
2009	784,222
2010	801,914
2011	823,910
2012	840,684
2013	885,226
2014	851,745
2015	844,433
2016	802,038
2017	774,238
2018	730,434

Visual Representation:



5. Future work

Currently, we are revisiting the original database to further analyze patients identified in our study. Our next step involves creating a comprehensive table that includes a patient's diagnostic codes along with all potentially relevant demographic information. This will allow for a detailed analysis of the cause-and-effect relationships in diabetes prevalence.

One challenge we are facing is the complexity of the data from 2016 to 2018. The database contains a significant amount of intricate information, which complicates the process of selecting and extracting the specific fields required for this analysis. To address this, we are running queries to combine the new tables with necessary information. While we have developed an algorithm for this process and potential future process, it remains time-consuming for running due to the large size of the database.

Once the data extraction is complete, the future analysis will focus on applying statistical and machine learning models to uncover key insights about the causes of diabetes which aim to improve understanding of demographic patterns and their impact on diabetes prevalence trends.