https://www.overleaf.com/project/6034854acdbf455955cd0ec2

PREVALENCE OF UNDIAGNOSED DIABETES TYPE 2 AND SOCIAL RISK FACTORS AMONG COMMUNITIES WITH DIAGNOSIS DISPARITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Type 2 Diabetes Mellitus (DM2) is a highly prevalent chronic condition which, when unmanaged, can negatively affect patient lives, along with health systems and payers, due to increased costs of care (Greenapple, 2011). Timely diagnosis is critical in managing this disease. We will use advanced analytics and a hybrid recommender system machine learning approach to assist a large mid-western health system in understanding their current undiagnosed DM2 patients, and patients likely to be classified with DM2 over the next 12 months. We will surface these results on an analytical dashboard that will identify social and health risk factors associated with the undiagnosed diabetic population. The health system will then develop effective strategies to manage the disease at the population level, such as ensuring increased adherence to evidence-based measures, rolling out digital wellness applications, and increasing specialist headcount in under-served and underprivileged areas.

1 BACKGROUND

Type 2 Diabetes mellitus (hereafter, DM2) is a highly prevalent chronic condition. Key to effective management of DM2 is timely diagnosis. According to the CDC, estimates from 2018 indicate that 21% of US adults with DM2 were undiagnosed (CDC, 2020). Without diagnosis, glycemic control strategies and other evidence-based therapies to delay progression of the disease cannot be instituted. This delay in diagnosis can be costly – to patients, in terms of health status and disease progression, livelihood and life expectancy, and to health systems and payers, due to the resulting increased incurred costs of care (Greenapple, 2011). In the US alone, 1 out of 7 dollars spent on medical care is related to Diabetes, and by 2025 it is expected that approximately 20 million people will have DM2 with 6 million of them undiagnosed (Black, 2002).

Minority groups, such as African Americans, or those belonging to a lower socioeconomic class, tend to have higher percentages of undiagnosed DM2 compared to non-Hispanic Whites. These groups are at risk for more complications due to unmanaged disease (Black, 2002). By the time symptoms appear in an undiagnosed DM2 patient, irreversible complications may have developed, such as neuropathy or retinopathy. Although family history is indicatory of DM2, other social factors play into the onset, management and exacerbations of the disease. Poor access to nutritious foods and physical activity, largely seen in low-income neighborhoods can double the odds of developing DM2 and can increase chances of complications from kidney problems to limb amputations (Presser, 2020).

A large Illinois-based health system has seen a sizeable gap between the prevalence of DM2 among patients as coded in their electronic medical record (EMR) versus the expected prevalence in their geographic area. For example, the Illinois Department of Public Health estimates the prevalence of DM2 in Illinois to be 12.5% with 1 in 4 of those with DM2 being undiagnosed (2020). The prevalence of DM2 within the health system's Illinois location is estimated to only be 4.2%. The health system also conducted an analysis to understand their South Asian population and the DM2 prevalence within that community. They learned that there was a 25.8% gap in the DM2 Diagnosis within that cohort alone (Sitafalwalla, 2020). This health system aims to apply Machine Learning

(ML) algorithms to their structured EMR data, regional Social Determinants of Health, and Geospatial datasets to understand other undiagnosed DM2 patient populations within their service areas. They intend to use ML outputs to understand social factors, such as race, that may create barriers to appropriate and timely healthcare for individuals (Heath, 2021). When algorithms incorporate race and ethnicities however, there is a risk of misinterpreting outputs and propagating inequities. There are many associations that can be made between a person's race and their clinical outcome; however the assumption of causality can lead to worse health outcomes for minority groups (Vyas et al., 2021). Adjusting ML outputs based on race will have to be taken into consideration when trying to understand diagnosis disparities of DM2 by social factors, such as race. With early confirmation of a DM2 diagnosis, multidisciplinary diabetes care teams can intervene quickly and implement measures to delay the progression of the disease.

2 ANTICIPATED METHOD

2.1 OBJECTIVE

The goal of our project is to use advanced analytics and machine learning to assist a large health system in understanding the diabetic population within their patient service areas. The team aims to understand how to best improve their delivery of care. We will develop an analytical dashboard that will provide insights on the healthcare system's diabetic patient population. We aim to identify the social risk factors associated with the undiagnosed diabetic population. The health system can use these insights to aid them in their community intervention decisions and development of effective strategies for effective DM2 management.

2.2 METHODS

2.2.1 COHORT SELECTION

The patient cohort will include adults (ages 18-75) receiving active care in the health system, defined as individuals who have had at least two ambulatory care visits in the past 24 months.

2.2.2 OUTCOME DEFINITION

We define the diagnosed DM2 cohort as individuals in our cohort who have documented DM2 based on diagnostic coding for DM2 via ICD-10-CM codes (in alignment with Electronic Clinical Quality Measures regarding DM). Individuals who do not meet these criteria would be considered as undiagnosed. We aim to identify patients who should be evaluated for DM2 and show signals of the disease in the data. Signals include, prescribed/current medications, lab results, family and social history, and social determinant data. We will exclude type 1 diabetics, pregnant women, those older than the age of 75, frailty and dementia patients (Peña, 2019), patients at end of life (e.g. Do Not Resuscitate/Do Not Intubate patients), and outreach laboratory-only patients (patients with only hospital serviced lab encounters).

2.2.3 DATA TRANSFORMATION

We will be using the following data sources onto an Azure Synapse instance (see citations in references section). These datasets will provide us with a breadth of location-specific knowledge to understand influential factors in missing a DM2 diagnoses.

Data source ingestion

Data Source	Description
Epic (EMR Data warehouse)	This will include patient encounter and demographic data. Patient diagnosis, procedure, insurance and medication history, provider details and lab values
American Community Survey (ACS)	These tables provide social, economic, housing and demographic data for a selected geography. We will be looking at Illinois and Wisconsin data by Zip Code
AHRQ's Social Determinants of Health	This dataset aggregates data from a collection of federal and publicly available datasets such as, CDC datasets, US Census, Area Health Resource Files, Social Vulnerability Index and more. We will only pull in variables that are related to our use case (we will run an exercise to understand relevant values).
NPI Registry	This dataset gives us provider details (location, specialty, association with Provider Group Practice and Provider-Hospital Organization)

2.2.4 FEATURE ENGINEERING

We will utilize features that will be defined by our data analysis, review of related literature, and input from our clinical informatics team. We will assess the importance of our features to model predictions using Shapley values and plots (Casas, 2019). A full list of features is still under development. Below are examples of categories and features we will incorporate in our models and/or descriptive dashboard:

Category	Example Features
Demographic	Age Group (<18, 18-34, 35-44, etc.), Sex, Insurance Type, Payer
Socioeconomic	Social Vulnerability Index in Census Tract, Count of Fast Food Restaurants in Zip, Employment Status
Cormobitities	HCC's, Diagnoses on prior encounters and claims, Procedures on prior encounters and claims, Lab Results
Utilization	Count of PCP Visits, Count of DM2 Specialist Visits, Count of Preventative Care Visits
Geospatial	Distance from PCP, Distance from Specialist

2.2.5 RECOMMENDER SYSTEM AND CLASSIFIER DEVELOPMENT

We will analyze the prevalence of undiagnosed DM2 via a hybrid recommender system approach. We will develop content-based filtering and collaborative filtering models to 'recommend' a DM2 diagnosis for undiagnosed patients who present similar qualities to those with a documented and known diagnosis. See design positive class assignment mechanism in Figure 1.

We will model the incidence of DM2 diagnosis by risk stratifying patients likely to be diagnosed with DM2 in the next 12 months. We are exploring two methods to train on our diagnosed cohort: a binary classification model or survival model that predicts a "time to" diagnosis event. We will also deploy our Fairness in ML Toolkit, since race and socioeconomic backgrounds of patients may be incorporated in our models, this will allow us to evaluate fairness and bias in our machine learning models. Our toolkit will generate comparison tables which will compare our models with fairness-aware alternate versions of those same models.

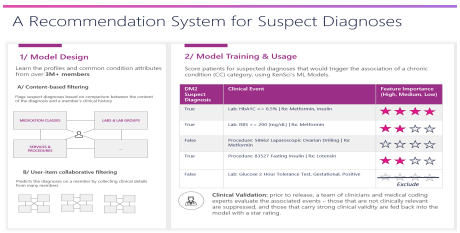

Figure 1: Model design

2.2.6 Model Output Evaluation

Once we have our model outputs, we will compute the proportion of DM2 observed within identified cohorts (e.g. Non-Hispanic White Women between the ages of 18-34 who are taking Metformin). Then with collaborative filtering, we will predict which patients will likely have a diagnosis of DM2 based on a diagnosis of DM2 among similar members. Afterwards, we will aggregate our findings to understand the geospatial distribution of the DM2 population.

2.2.7 User Interface Design

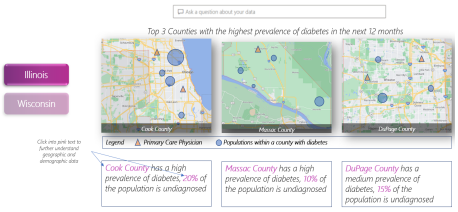

Figure 2: Illustrative mock-up of a narrative analytic dashboard.

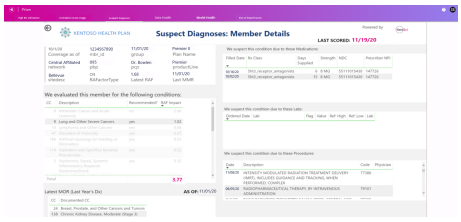

Figure 3: Suspect Diagnosis dashboard, patient drill down.

Our model results and exploratory analysis will be surfaced on a dashboard for the health system. These healthcare leaders will be interested in an executive level dashboard (Figure 2), where the prevalence of DM2 in relation to a geographic location and its population is represented in a narrative format. It will also include the percentage of patients who were not formally diagnosed with DM2 and what their demographic features are, to help healthcare leaders understand what types of patient cohorts slip through the cracks and where it is the most common. This interface will allow healthcare leaders to make informative decisions in order to implement programs for early DM2 detection and patient care management.

2.2.8 Results

We are currently in the beginning stages of this work and do not have results at this time.

3 REFERENCES

American Community Survey, 2019 American Community Survey 5-Year Estimates, U.S. Census Bureau, www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/.

Austin, Steven R, et al. "Why Summary Comorbidity Measures Such As the Charlson Comorbidity Index and Elixhauser Score Work." Medical Care, U.S. National Library of Medicine, Sept. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC3818341/.

Black, Sandra. "Diabetes, Diversity, and Disparity: What Do We Do with the Evidence?" American Journal of Public Health, © American Journal of Public Health 2002, Apr. 2002, www.ncbi.nlm.nih.gov/pmc/articles/PMC1447113/.

Casas, Pablo. "How to Interpret SHAP Values in R (with Code Example!)." Data Science Heroes Blog, Data Science Heroes Blog, 19 Mar. 2019, blog.datascienceheroes.com/how-to-interpret-shap-values-in-r/.

Diabetes. Illinois Department of Public Health, 2020, https://dph.illinois.gov/topics-services/diseases-and-conditions/Diabetes

Greenapple, Rhonda. "Review of Strategies to Enhance Outcomes for Patients with Type 2 Diabetes: Payers' Perspective." American Health amp; Drug Benefits, Engage Healthcare Communications, LLC, Sept. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC4105736/.

HCUP CCS. Healthcare Cost and Utilization Project (HCUP). March 2017. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Heath, Sara. "Top Social Determinants of Health Barring Patient Care Access." PatientEngagementHIT, 28 Jan. 2021, https://patientengagementhit.com/news/top-social-determinants-of-health-barring-patient-care-access

NPPES, Center for Medicare and Medicaid Services, 2 Feb. 2021, www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/DataDissemination. Pastors, Joyce Green, et al. "The Evidence for the Effectiveness of Medical Nutrition Therapy in Diabetes Management." Diabetes Care, American Diabetes Association, 1 Mar. 2002, https://www.care.Diabetesjournals.org/content/25/3/608.

Peña, Cindy. "Improving Care for Those with Advanced Illness and Frailty." NCQA Blog, 27 Mar. 2019, blog.ncqa.org/improving-care-advanced-illness-frailty/.

Presser, Lizzie. "The Black American Amputation Epidemic." ProPublica, 19 May 2020, features.propublica.org/Diabetes-amputations/black-american-amputation-epidemic/.

"Prevalence of Both Diagnosed and Undiagnosed Diabetes." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 24 June 2020, www.cdc.gov/Diabetes/data/statistics-report/diagnosed-undiagnosed-Diabetes.html.

Schmittdiel, Julie A, et al. "Population Health Management for Diabetes: Health Care System-Level Approaches for Improving Quality and Addressing Disparities." Current Diabetes Reports, U.S. National Library of Medicine, May 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5536329/.

Shoeb J. Sitafalwalla. Opportunities to Increase Awareness through Social Media: A case study in population health innovation. National Lipid Association Scientific Session: Miami, FL, May 17th, 2019

Social Determinants of Health Database (Beta Version). Content last reviewed December 2020. Agency for Healthcare Research and Quality, Rockville, MD. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html

Vyas, Darshali A., et al. "Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms: NEJM." New England Journal of Medicine, 17 Feb. 2021, www.nejm.org/doi/full/10.1056/NEJMms2004740.