Prediction of type 2 diabetes in the US: an analysis and application of nationally representative data from National Health and Nutrition Examination Survey

Data analysis

Naiwen Ji (nj294), Yuze Qin (yq276), Zhaohan Xing (zx333)

ORIE 5741

Problem overview:

The interested prediction question is:

With general information related to one's socio-demographic, lifestyle and non-invasive anthropometric characteristics, can we predict one's diabetes status?

Type 2 diabetes is the most common type of diabetes in the world, consistent with 95% of diabetes cases. In the US, about 38 million people have diabetes (1). Diabetes alone is the eighth leading cause of death in the US. However, according to a recent study, about 34.3 % of the diabetic population were unaware of their diabetes condition (2). In the past two decades, the prevalence of type 2 diabetes increased dramatically in the US. In 2020, the estimated prevalence of prediabetes was 38.4 million people aged over 18 years had prediabetes which consisted of about 12% of the total US adult population (3). Type 2 diabetes is also an important risk factor for various medical conditions and chronic diseases including blindness, kidney failure, heart attack, stroke, and lower limb amputation. Studies also showed that diabetes was associated with an increased risk of most cancers and infectious diseases including COVID-19 (4).

The American Diabetes Association (ADA) published an economic cost of Diabetes in the US in 2022 which indicated that the total annual cost of diabetes was 412.9 billion US dollars including 306.6 billion in direct medical costs and 106.3 billion in indirect costs. The indirect costs of diabetes were related to disability, presenteeism, and lost productivity due to premature death resulting in diabetes. Although diabetes as a chronic disease is manageable and has relatively slower progression compared to infectious disease, as shown in the report, it is causing additional loss of revenue and extra expenses for health insurance (5). The standards of care for people living with diabetes include diagnosis and classification of diabetes, glycemic monitoring and management, comprehensive medical evaluation and assessment of comorbidities, and facilitating positive health behaviors and well-being to improve health outcomes (6). The health insurance industry is highly involved in these processes of diabetes management. To better prevent diabetes-related economic loss and the decision-making process of health insurance, there is an urgent need to understand the dynamics of diabetes and develop a prediction model for diabetes.

Machine learning-induced prediction model is a rising method in the public health and nutrition fields. A good example of a successful implementation of the algorithm in the prediction of cardiovascular diseases (CVD) was the ASCVD risk estimator conducted and implemented by the American College of Cardiology. The prediction model was developed based on the generalized linear regression model and the model is widely used in research and clinical settings as an assessment and prevention tool (7–9). As for diabetes, there is no existing commonly used prediction model. However, in a 2019 literature review, the authors mentioned the potential opportunities of utilizing machine learning (ML) classifier algorithms to develop prediction

models for diabetes (10). Studies have compared the performance of ML classification algorithms such as logistic regression, forests, SVM, artificial neural networks, and Naive Bayes to predict diabetes. Both of the studies achieved decent accuracy which indicates the feasibility of this proposed project (11,12). With access to the most updated nationally representative dataset in the US with high-quality data, we will be able to develop a prediction model for diabetes.

Dataset availability

Data will be utilized in this study from the National Health and Nutrition Examination Survey (NHANES). Survey design and details about data collection were available elsewhere (13). Briefly, NHANES is a publicly available, large-scale, and nationally representative health-related survey conducted in the US. In this study, we will use data from NHANES cycle 2017-2018. Participants aged over 18 years with oral glucose tolerance test and/or HbA1c measurements will be included in the analyses.

From previous literature review, the potential correlates of diabetes including socio-demographic information including age, sex, race/ethnicity (non-Hispanic White, non-Hispanic black, Hispanic, and other), education level (less than a high school diploma, high school graduate/GED, some college/AA degree, and college graduate or more)(14), and household income (≤130% (reference group), >130% to 350%, and >350% by the ratio of family income to poverty (FPL)) (15) will be included in this study. Lifestyle factors including smoking (none, former, current smoker) (16), alcohol consumption (none, light, moderate, heavy drinker) (17), diet and physical activity level (having met (≥600 MET-minutes/week equivalent to 150 min/week of moderate intensity or 75 min/week vigorous-intensity physical activity) or not meet (<600 MET-minutes/week)) (14) will also be included in the analyses. Body mass index (BMI) and body fat composition will be included in the analyses as well.

Our Approaches

In this cross-sectional study, the descriptive of each included feature will be estimated adjusting for sample weight and reported as n% (95% CI). For ML approaches, the output will be diabetes as defined by Hb1Ac. The output will be encoded as a binary dummy variable. Due to the nature of the algorithm suitable for classification, all categorical features will be encoded as dummy values, and the continuous variables will be normalized before input in the ML models. The dataset will be split to two, training and testing datasets, by the ratios of 80/20. Feature selection will be conducted by comparing different methods including principal component analysis (PCA), lasso regression and correlation coefficient selection based on biological plausibility. Feature engineering methods will be used to optimize the performance of ML models. Different supervised ML algorithms will be used including support vector machine (SVM), random forest, logistic regression and k-nearest neighbors (KNN).

All models mentioned above will be tested for performance. The evaluation will be based on sensitivity, specificity, accuracy and likelihood ratio. Area under receiver operating characteristic curve (ROC) will be used to assess the performance of the models. The cross-validation procedure will be used to ensure the performance and to preserve generalizable model conclusions. A 10-fold cross-validation will be used in the training set. Then the best-fit model will be tested to predict diabetes in the testing set we previously defined.

Reference

1. Diabetes Quick Facts | Basics | Diabetes | CDC [Internet]. 2023 [cited 2024 Mar 12]. Available from: https://www.cdc.gov/diabetes/basics/quick-facts.html

2. Menke A, Casagrande S, Avilés-Santa ML, Cowie CC. Factors Associated With Being Unaware of Having Diabetes. Diabetes Care. 2017 Mar 13;40(5):e55–6.

3. National Diabetes Statistics Report | Diabetes | CDC [Internet]. 2023 [cited 2024 Mar 12]. Available from: https://www.cdc.gov/diabetes/data/statistics-report/index.html

4. NATIONAL CENTER FOR FARMWORKER HEALTH [Internet]. [cited 2024 Mar 12]. Diabetes Fact Sheet. Available from: https://www.ncfh.org/diabetes-fact-sheet.html

5. Parker ED, Lin J, Mahoney T, Ume N, Yang G, Gabbay RA, et al. Economic Costs of Diabetes in the U.S. in 2022. Diabetes Care. 2024 Jan 1;47(1):26–43.

6. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. Introduction and Methodology: Standards of Care in Diabetes—2023. Diabetes Care. 2022 Dec 12;46(Supplement_1):S1–4.

7. Ponce de León-Ballesteros G, Sánchez-Aguilar H, Aguilar-Salinas CA, Herrera MF. Improvement of the 10-Year Atherosclerotic Cardiovascular Disease (ASCVD) Risk Following Bariatric Surgery. Obes Surg. 2020 Oct;30(10):3997–4003.

8. Vassy JL, Posner DC, Ho YL, Gagnon DR, Galloway A, Tanukonda V, et al. Cardiovascular Disease Risk Assessment Using Traditional Risk Factors and Polygenic Risk Scores in the Million Veteran Program. JAMA Cardiol. 2023 Jun 1;8(6):564–74.

9. Ten-year risk assessment for cardiovascular diseases using ASCVD risk estimator plus: outcomes from hypertension and diabetes patients - PMC [Internet]. [cited 2024 Mar 12]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10604522/

10. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences. 2019 Jan;9(21):4604.

11. Hasan MdK, Alam MdA, Das D, Hossain E, Hasan M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access. 2020;8:76516–31.

12. Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Computer Science. 2018 Jan 1;132:1578–85.

13. NHANES - National Health and Nutrition Examination Survey Homepage [Internet]. 2023 [cited 2023 Nov 7]. Available from: https://www.cdc.gov/nchs/nhanes/index.htm

14. Scholes S, Bann D. Education-related disparities in reported physical activity during leisure-time, active transportation, and work among US adults: repeated cross-sectional analysis from the National Health and Nutrition Examination Surveys, 2007 to 2016. BMC Public Health. 2018 Jul 28;18:926.

15. Wang K, Zhao Y, Nie J, Xu H, Yu C, Wang S. Higher HEI-2015 Score Is Associated with Reduced Risk of Depression: Result from NHANES 2005–2016. Nutrients. 2021 Jan 25;13(2):348.

16. ALHarthi SSY, Natto ZS, Midle JB, Gyurko R, O'Neill R, Steffensen B. Association between time since quitting smoking and periodontitis in former smokers in the National Health and Nutrition Examination Surveys (NHANES) 2009 to 2012. J Periodontol. 2019 Jan;90(1):16–25.

17. Gay IC, Tran DT, Paquette DW. Alcohol intake and periodontitis in adults aged ≥30 years: NHANES 2009–2012. Journal of Periodontology. 2018;89(6):625–34.