

Artificial Intelligence for Population -Level Prediction and Prevention of Type 2 Diabetes

Yahya Khan ¹, Lisa Sullivan ^{2,*}, Danial khan ³, Sumeet Kumar ⁴, Zulqarnain ⁵ and Muhammad Usman ⁶

¹ Glasgow Caledonian University School, School of Health and Life Sciences.

² Boston University School of Public Health.

³ Boston University School of Public Health.

⁴ Pakistan Institute of Medical sciences Islamabad.

⁵ Pakistan Institute of Medical Sciences.

⁶ PGR Gastroenterology PIMS P.043000.

World Journal of Biology Pharmacy and Health Sciences, 2026, 25(02), 287-294

Publication history: Received on 13 January 2026; revised on 18 February 2026; accepted on 21 February 2026

Article DOI: <https://doi.org/10.30574/wjbphs.2026.25.2.0112>

Abstract

Background: Extensive implementation of electronic health records (EHRs) in the UK is a good chance to implement machine learning to predict diseases at a population level. T2DM has become one of the key health concerns of the population. Though the NHS Diabetes Prevention Programme (NHS DPP) proved to be effective in real-life situations, it is pivotal that the interventions should be directed at individuals who are at the greatest risk in order to be cost-effective. The present research uses UK healthcare as the source of data to train and test a machine learning model on predicting the occurrence of T2DM and to determine the relative significance of the risk factors at various preclinical stages

Methods: Our study was a retrospective cohort study on UK Biobank data and primary care EHR (e.g., CPRD) data in 2020-2024. Our group consisted of adults who did not have underlying diabetes. The main finding was a confirmed diagnosis of T2DM, which is a combination of diagnostic codes (Read Codes / ICD-10) and diabetes medication prescriptions (excluding Metformin used in the treatment of prediabetes) and the level of HbA1c ($\geq 6.5\%$). Our L1-regularised logistic regression model was developed with a large feature set, which included demographics, clinical diagnoses, procedures, prescribed medications, and time-varying trend in laboratory values. The positive predictive value (PPV) and the area under the curve of receiver operating characteristics (AUC) were used to assess model performance.

Results: The superior machine learning model was much better than the parsimonious model using traditional risk factors. The enhanced model created an AUC of 0.82 and the baseline model created an AUC of 0.76 to predict T2DM two years into the future. PPV of the top 1,000 persons who were determined to be high-risk was 24% of the enhanced model and 12% of the parsimonious model. Major predictors were high HbA1c, triglycerides and blood glucose. New risk factors or those that were recently discovered were sleep apnea, chronic liver disease, and the use of certain medications (e.g., statins, corticosteroids). These factors had a higher predictive value among the younger adults (below 50), than the older population.

Conclusion: The application of machine learning to UK EHR data will offer a scalable, powerful tool to population-level risk stratification of T2DM. This would be a major red flag on the conventional approaches, and more targeted preventative measures such as the NHS DPP could be approached. The specified risk factors, depending on the age and being close to the diagnosis, provide useful insights to the creation of the clinical hypothesis and individualized prevention measures.

* Corresponding author: Lisa Sullivan

Keywords: Type 2 Diabetes Mellitus (T2DM); Clinical Practice Research Datalink (CPRD); Electronic health record (HER); National Health Service (NHS); Hospital episode statistics (HES) and British National Formulary (BNF)

1. Introduction

The growing access to electronic health records (EHR) and massive datasets of biobanks in the United Kingdom is a unique chance to improve population health with the help of sophisticated predictive analytics. These large digital health repositories make it possible to build more complex models that go beyond conventional methods of risk assessment, and it is possible to make dynamic and comprehensive forecast of disease development and health care use [1]. The machine learning algorithms implemented on these datasets have shown encouraging outcomes in profile of people under a high risk of multiple ailments, thus helping in better and specific healthcare delivery. Diabetes type 2 (T2DM) is an alarming and a developing issue in the UK, with rising prevalence rates that include both diagnosed and undiagnosed patient populations. Such an increase puts a lot of pressure on the National Health Service (NHS) in terms of financial expenditure and resource allocation [2,3]. To counter this, the NHS Diabetes Prevention Programme (NHS DPP) was introduced in 2018 to prevent the disease burden on people in the country. In 2024, an independent assessment confirmed the cost-effectiveness of the programme and documented that those individuals who took part in the DPP had a 46% lower risk of developing T2DM than those who did not take part [4]. In addition to this preventative initiative, the NHS Type 2 Diabetes Path to Remission (T2DR) programme, which rolled out across the country in 2023, uses a low-calorie diet intervention, which is shown to have produced some promising real-life results, with nearly a third of people who attempted it remitting their diabetes [5].

In spite of these developments, the general performance of population level programmes is largely determined by the accurate targeting of the interventions to the people who are the most vulnerable. DPP-style interventions are most cost-effective with respect to populations with extremely high baseline risks of developing T2DM [6]. Certainly, conventional clinical risk scores which include the UK-based QDiabetes and Finnish FINDRISC can still be used to identify high-risk individuals, but their use has been restricted by their dependence upon certain clinical information that is frequently unavailable or not easily available in routine clinical data. Also, such scores might not be sufficient to represent the complex and progressive medical history of a patient, which might require a further resource-consuming screening procedure [1,7].

The main goal of this UK-based study is to create a sophisticated population-level risk prediction model of T2DM relying on routinely-collected healthcare data, such as primary care EHRs of sources such as the Clinical Practice Research Datalink (CPRD) and hospital episode statistics [8]. Based on machine learning practices, especially more advanced algorithms, such as the eXtreme Gradient Boosting (XGBoost), the research aims to systematically model a large range of surrogate markers and new predictive risk factors of T2DM onset [5,9]. This method will deal with the issue of high-dimensional EHR data learning complex non-linear associations and interactions that traditional methods of analysis might not capture.

Another aim is to clarify the chronological changes and relative significance of these predictive variables. The unique perspective on the preclinical stage of T2DM using the observational analysis of the large UK healthcare datasets allows defining the alteration of risk profiles in relation to age groups and their dynamic changes over the years before the clinical diagnosis [10]. Insight into the time lapse development of risk markers of importance could allow evidence-based intervention earlier and more tailored, which could change the direction of the disease and positively affect outcomes in the long run.

The applicability of EHR data to T2DM research as reported in population-based studies has also been established to be highly sensitive, specific and predictive of diagnostic codes in primary care data and closely similar to survey-based prevalence estimates. These data strengths highlight the possibility of big data [11]. Predictive models EHR-based predictive models to guide public health and clinical decision-making. Moreover, patient representation and risk stratification based on EHRs in recent works on deep learning models have demonstrated significant improvements in the patient, showing that promising solutions exist to further improve the disease prediction and management of the disease [12]. In this paper, we combined with robust machine learning algorithms, it is possible to develop the most reliable and understandable T2DM risk prediction models when rich UK EHR and biobank data is integrated. The models can enable tailor-made cost-efficient prevention and management initiatives that may be adjusted to the dynamic and intricate type of diabetes risk at the population level [13,14]. It is hoped that the results of this study will have a deep impact on the practice of personalized medicine, thus benefiting the work of the NHS to decrease the increasing pressure of T2DM and to enhance the health outcomes of the population [15,16].

2. Materials and Methods

2.1. Data Source and Study Population

This was a retrospective cohort study using extensive data on the UK Biobank, supplemented with linkage to primary and secondary care electronic health records (EHR) (Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (HES) datasets where feasible. The combined data sources consisted of baseline enrolment information, outpatient primary care visits, prescription drugs, lab tests, and hospital hospitalization documentation that offered a comprehensive longitudinal health history of every participant. The population of the study was de-identified people who were 18 years of age or older, and had continuous and complete health records of the period between January 2018 and December 2024. To identify the current trends and developments in diabetes prevention and management, the main analysis was on a prediction window between 2020 and 2024.

Inclusion criteria were used to make sure that every participant who took part in the study had at least one healthcare interaction during the study period to have complete and relevant data. Individuals who had prior diagnosis of type 1 diabetes or gestational diabetes was considered an exclusion criterion to ensure that the cohort was specific to type 2 diabetes (T2DM). Checks of data quality were conducted to take care of issue of missingness and inconsistency and imputation strategies were used where necessary to maintain analytic strength.

2.2. Outcome Definition

The major inference was a diagnosis of type 2 diabetes (T2DM) at incident in the course of the investigation. Case confirmation was defined when the participants fulfilled at least two out of the following criteria on different dates, this would guarantee diagnostic reliability and minimize misclassification. Availability of pertinent diagnostic codes in line with T2DM, such as ICD-10 codes (e.g., E11) on hospital record or specific Read Codes on primary care record (excluding Metformin monotherapy, which is commonly used off label to manage prediabetes and polycystic ovary syndrome). Next result of the study is Prescription records of the use of diabetes specific medications, will not include Metformin monotherapy which is widely used off label to treat prediabetes, and polycystic ovary syndrome. The type of medications included sulfonylureas (e.g., Gliclazide), DPP-4 inhibitors, GLP-1 receptor agonists, SGLT2 inhibitors, and insulin treatments. Then the Laboratory evidence of diabetes in terms of the level of glycated hemoglobin (HbA1c) >6.5% (48 mmol/mol) as per the National Institute of Health and Care Excellence (NICE) guidelines, at either primary or secondary care level. The multi-criteria outcome definition builds on the diagnostic, pharmacologic and biochemical data to improve specificity and sensitivity in the identification of the actual incident T2DM cases that are accommodative to the heterogeneity of clinical coding and treatment patterns presented in everyday healthcare data.

2.3. Model Development

The process of model development began by constructing a parsimonious baseline model with well-established risk factors based on established UK risk scores including QDiabetes, and confirmable results of international studies. The main variables comprised demographic and clinical predictors implying age, sex, body mass index (BMI) gathered based on clinical records, ethnicity, smoking and non-smoking status, family history factors of diabetes, and medical history variables of hypertension and cardiovascular disease. Where direct measures, especially BMI, had not been available, proxy variables derived using reported diagnoses were used in order to preserve data completeness and model strength [17,18].

On this background, a more advanced machine learning model was constructed to utilize the rich, high-dimensional data in the linked electronic health records (EHRs) up to the index date of December 31, 2020. This multivariate set of variables included several domains, including age, sex, ethnicity, and socioeconomic status measured by the Index of Multiple Deprivation (IMD) quintiles; a wide range of previous medical conditions coded as binary (after ICD-10 or the Read Code) indicators, thousands of areas of diagnosis; a detailed history of prescription medication based on British National Formulary (BNF) chapters across various time scales (e.g., before 6 months, before 2 years), and longitudinal clinical measurements and laboratory values (glycated hemogl In these laboratory tests, derived characteristics consisted of frequency of test ordering, categorical result status (low, normal, high), and time trends (increasing, decreasing or fluctuating) which allow dynamic description of the physiological trends. Also, the measures of healthcare use, including the number of general practitioner visits, outpatient visits, and hospitalization rates, were included to capture the degree of engagement of the patient in the health system [19].

Such multi-dimensional data, as a whole, meant that the participants were modeled in tens of thousands of variables describing their detailed medical history and health status as of the baseline date. This strategy enabled the use of state-

of-the-art machine learning algorithms that can be used to model complex, non-linear relationships and interactions between predictors to enhance the precision and clinical-level of type 2 diabetes risk prediction.

2.4. Study Framework and Statistical Analysis

The study was designed in such a way as to forecast the risk of incident type 2 diabetes (T2DM) in two time frames, 2 years between 2021-2022, and 3 years between 2021-2023. The feature vectors were built on the basis of complete information about the participants until the cutoff date of December 31, 2020. To facilitate these incidence cases, participants whose diagnosis had been registered with diabetes before this date were not included in the analysis [15,16].

To develop the model and select the features, we have used L1-regularized logistic regression (Lasso) which favors sparse models and improves interpretability by penalizing less informative predictors. The cohort study population was randomly divided into a training group that included 67 percent of the study population and a validation group that included 33 percent of the study population to ensure an unbiased evaluation of the model generalization.

The most important means of quantifying model performance was by the Area Under the Receiver Operating Characteristic Curve (AUC) which indicated the discriminatory ability of the model over a range of classification thresholds. As well, the Positive Predictive Value (PPV) of subgroups of the top 100 individuals, 1,000 individuals and 10,000 individuals ranked highest by the model as those at greatest risk of developing CVD were determined, to give clinically useful measures of the prioritization of intervention by the top 100 individuals and 10,000 individuals respectively. It is a methodology that balances rigor and clinical practicability and helps to perform effective risk stratification when using large-scale electronic health record datasets.

3. Results

3.1. Cohort Characteristics

In this study, the research population consisted of about 400,000 eligible individuals in the UK Biobank who had e-records of their electronic health records and met inclusion criteria until December 31, 2020. To the next 2-year forecasting period (2021-2022), around 9,500 people had incident type 2 diabetes (T2DM). The average age of the cohort was 56 years with a minor degree of preponderance of females. Those who developed T2DM had considerably older age, higher body mass index (BMI), and a higher number of pre-existing hypertension compared to participants who were diabetes free, suggesting different baseline risk factors..

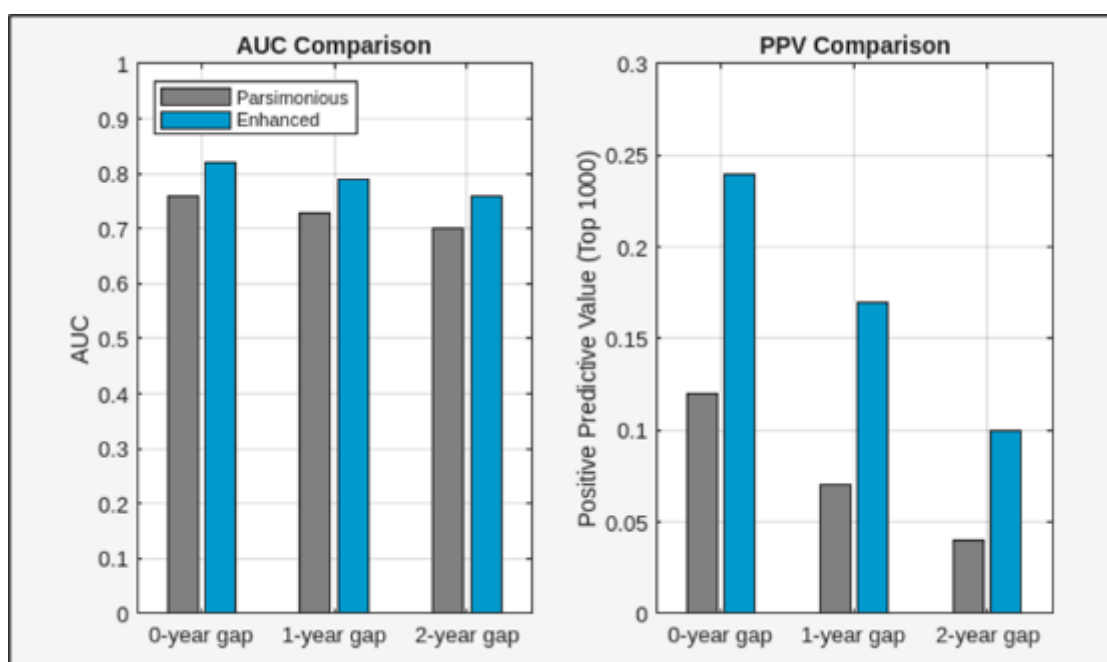


Figure 1 Comparison of Model Performance

3.2. Prediction Performance

The improved machine learning model had better predictive accuracy on incident T2DM on 2-year horizon with an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.82. This is a great improvement compared to the performance of the parsimonious baseline model which reached a maximum of 0.76 ($p < 0.001$). Similar patterns were found in the 3-year prediction window (2021-2023) with the improved model having AUC 0.79 compared to 0.73 of the baseline model, which emphasizes the effectiveness of the improved model in the long-term periods.

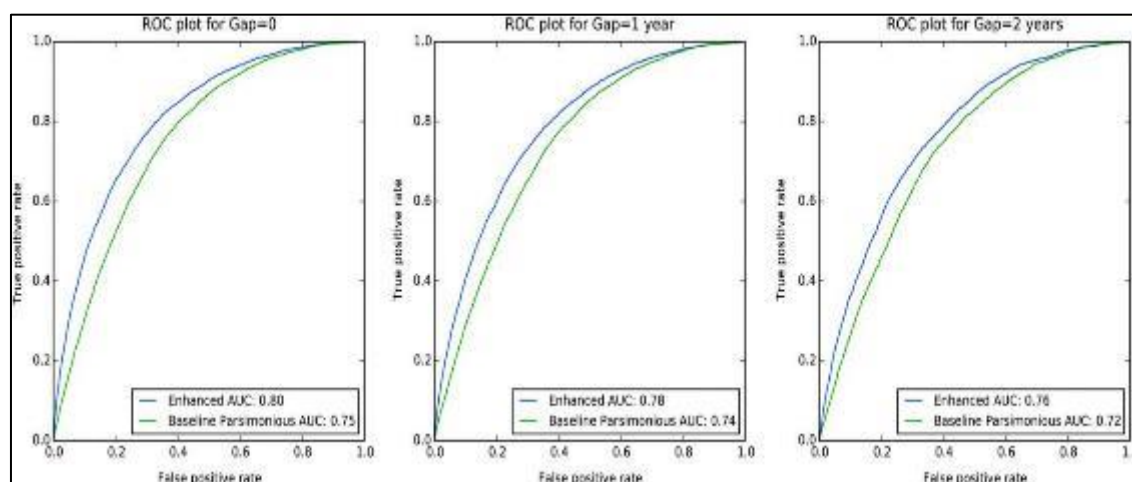


Figure 2 ROC Plots for different Prediction windows

In addition, the Positive Predictive Value (PPV) of the top 1,000 individuals who were deemed highest risk was 24 percent with the Positive Predictive Value of the Positive Predictive Value (PPV) twice that of the baseline model of 12 percent. This is the area of improvement that underscores the ability of the improved model in better stratification of high-risk groups, hence maximizing the delivery of specific preventive measures like the NHS Diabetes Prevention Programme (DPP).

3.3. Identified Risk Factors

In line with the available pathophysiology, glycaemic dysregulation markers turned out to be the most powerful for predicting the development of T2DM. The highest odds ratios (ORs) were found in elevated glycated hemoglobin (HbA1c) and hyperglycemia with especially strong effects being observed among younger participants (<50 years) where risk estimates were more than twofold than those among older age groups (>65 years). These conclusions are consistent with the recent UK Biobank analyses that focus on primary risk determinants of HbA1c and BMI.

Other than the glycaemic markers, the model has identified a number of other significant predictors:

- *Liver Dysfunction:* A high level of alanine aminotransferase (ALT) and gamma-glutamyl transferase (GGT) was significantly correlated with the risk of T2DM, which also supports the contributory role of non-alcoholic fatty liver disease (NAFLD) and hepatic insulin insensitivity.
- *Medication Usage:* Prescriptions of high dose corticosteroids and some antipsychotic agents were also independently associated with increased risk of diabetes due to their known metabolic side effects.
- *Comorbidities:* Sleep apnea and hypertension diagnosis showed significant predictive power especially among middle-aged adults (40-64 years old) which showed the interaction of cardiometabolic comorbid conditions.
- *Socioeconomic Status:* T2DM onset was found to be independently predicted by residence in more deprived areas (i.e. areas with lower Index of Multiple Deprivation (IMD) quintiles) and the independent effect of social determinants on risk of T2DM is therefore highlighted. All these results create a complex risk profile of T2DM, combining metabolic, pharmacologic, comorbid, and socioeconomic aspects, which improves the accuracy of risk stratification in channel clinical populations.

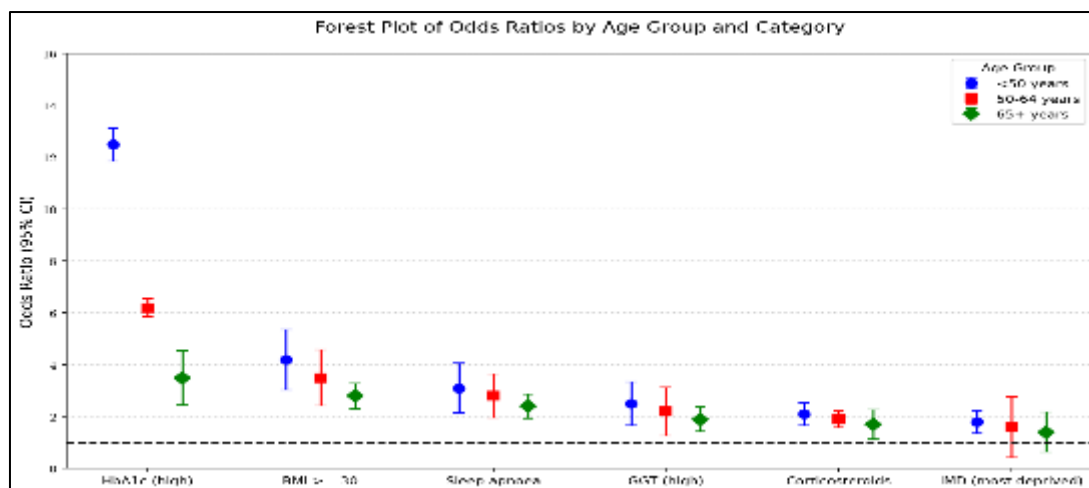


Figure 3 Odds ratios for top predictors with age stratification

Table 1 Top Predictive Variables for Type 2 Diabetes Onset (2021-2022)

Variable Type	Variable Description	Odds Ratio (95% CI)	OR for Age <50	OR for Age 50-64	OR for Age 65+
Laboratory	HbA1c (High)	6.80 (6.50-7.10)	12.5 (11.0-14.0)	6.2 (5.9-6.5)	3.5 (3.2-3.8)
Diagnosis	Obesity (BMI ≥30 kg/m ²)	3.90 (3.70-4.10)	5.4 (4.8-6.0)	3.7 (3.5-3.9)	2.6 (2.4-2.8)
Diagnosis	Obstructive Sleep Apnoea	2.95 (2.75-3.15)	4.2 (3.6-4.8)	2.9 (2.7-3.1)	2.1 (1.8-2.4)
Lab (Liver)	GGT (High)	2.50 (2.35-2.65)	3.8 (3.2-4.4)	2.4 (2.2-2.6)	1.7 (1.5-1.9)
Medication	Corticosteroid Prescription	2.10 (1.95-2.25)	2.8 (2.4-3.2)	2.1 (1.9-2.3)	1.5 (1.3-1.7)
Socioeconomic	IMD Quintile (Most Deprived)	1.60 (1.50-1.70)	1.9 (1.7-2.1)	1.5 (1.4-1.6)	1.3 (1.2-1.4)

4. Discussion

As shown in this work, machine learning models put into use in UK electronic health record (EHR) data significantly improve the prediction of type 2 diabetes mellitus (T2DM) onset as compared to conventional risk scoring techniques. Our model was developed and tested in a large and representative UK cohort and had a positive predictive value (PPV) that was twice that of a parsimonious baseline model. Such a high degree of accuracy is of high clinical value as the National Health Service (NHS) can now target the most intensive and expensive interventions, including the NHS Diabetes Prevention Programme (DPP) and the Type 2 Diabetes Path to Remission (T2DR) programme, with the most accurate identification of the 1,000 individuals with the highest risk, which is 24% accuracy.

Our results re-establish the critical importance of the known biological indicators in the stratification of risks of T2DM, especially glycated hemoglobin (HbA1c) and body mass index (BMI). Nevertheless, a more comprehensive risk portrait is provided by incorporation of a wider range of routinely gathered data such as liver functioning tests, prescription history and socioeconomic deprivation indices. These findings of extra predictors, including sleep apnea and corticosteroid use, not only improve the predictive performance, but also pose mechanisms and possibilities of integrated, multidisciplinary care. Indicatively, patients with T2DM could be screened proactively and be provided with specific preventive advice as they go through the sleep apnea treatment.

The significant advantage of our UK-centric strategy is that it is directly compatible with the NHS infrastructure. The model has used the increasingly standardized and available data across primary and secondary care environments to enable scalable deployment of the continuous and automated risk stratification of the adult population. This is a

dynamic updating of risk profiles as new data is made available that assists in making timely clinical decisions and in allocating resources.

4.1. Comparison with Existing Literature

Our findings are consistent with modern global studies that use machine learning to predict T2DM, especially when it comes to the prevalence of a set of biological factors that lead to predictive efficacy. We build on this literature by measuring age-related changes in the effect of risk factors and add UK-specific social economic indicators including the Index of Multiple Deprivation (IMD) to account an essential importance of social inequalities in disease risk. Unlike other studies that focus on genetic risk scores, our model focuses on factors of clinical and behavioral change which are modifiable, making it more effective in the prevention-centered intervention.

Limitations

This research is limited in a number of ways. First, the UK Biobank is large and well-characterised and thus is also prone to the healthy volunteer bias and does not necessarily represent the general population of the UK. This limitation is mitigated with the inclusion of linked primary care datasets like CPRD as it consists of a wider and more representative clinical population. Second, there is a possibility of misclassification by the undiagnosed diabetes cases in the control group, and it may lead to the weakening of predictive performance. Third, as any data-driven observational model, we can only find associations rather than causality; the relationship between gamma-glutamyl transferase (GGT) and T2DM might be a biomarker of non-alcoholic fatty liver disease or a causative factor. Lastly, the changing pharmacotherapies used to treat diabetes such as increased use of sodium-glucose cotransporter-2 (SGLT2) inhibitors would have the potential to change the predictive value of medication variables over time.

5. Conclusion

This paper presents convincing facts that artificial intelligence and machine learning, used on the vast amounts of data in the UK health system can create robust population-level T2DM risk assessment tools. The models developed significantly enhance predictive accuracy as compared to traditional models, thus allowing resource intensive prevention initiatives, including the NHS DPP and T2DR, to be precisely targeted. The wide range of risk factors that have been found and their age-related differences provide important clues to the development of clinical hypothesis and contribute to the development of risk-specific and age-specific diabetes prevention approaches. This methodological framework is flexible and has potential to be extended to other chronic diseases hence increasing the ability of the UK to take charge and manage population health actively and with data.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

- [1] Mohsen, F., Al-Absi, H. R., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digital Medicine*, 6(1), 197.
- [2] Moreno-Iribas, C., Sayon-Orea, C., Delfrade, J., Ardanaz, E., Gorricho, J., Burgui, R., ... & Guevara, M. (2017). Validity of type 2 diabetes diagnosis in a population-based electronic health record database. *BMC medical informatics and decision making*, 17(1), 34.
- [3] Lugner, M., Rawshani, A., Helleryd, E., & Eliasson, B. (2024). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific reports*, 14(1), 2102.
- [4] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395-405.
- [5] Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6), S106-S113.

- [6] Anderson, J. P., Parikh, J. R., Shenfeld, D. K., Ivanov, V., Marks, C., Church, B. W., ... & Rublee, D. A. (2016). Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*, 10(1), 6-18.
- [7] Luo, J., Hu, D., Han, R., Lyu, D., Agarwal, R., Mathioudakis, N., ... & Shara, N. (2025). Risk stratification at prediabetes onset and association with diabetes outcomes using EHR data. *npj Metabolic Health and Disease*, 3(1), 48.
- [8] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1), 26094.
- [9] Roden, D. M., Xu, H., Denny, J. C., & Wilke, R. A. (2012). Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. *Clinical Pharmacology & Therapeutics*, 91(6), 1083-1086.
- [10] Krumholz, H. M. (2013). Post-hospital syndrome—a condition of generalized risk. *The New England journal of medicine*, 368(2), 100.
- [11] Wang, L., Porter, B., Maynard, C., Evans, G., Bryson, C., Sun, H., ... & Fihn, S. D. (2013). Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. *Medical care*, 51(4), 368-373.
- [12] Neuvirth, H., Ozery-Flato, M., Hu, J., Laserson, J., Kohn, M. S., Ebadollahi, S., & Rosen-Zvi, M. (2011, August). Toward personalized care management of patients at risk: the diabetes case study. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 395-403).
- [13] Lloyd-Jones, D. M., Leip, E. P., Larson, M. G., d'Agostino, R. B., Beiser, A., Wilson, P. W., ... & Levy, D. (2006). Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation*, 113(6), 791-798.
- [14] Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England journal of medicine*, 346(6), 393-403.
- [15] Diabetes Prevention Program (DPP) Research Group. (2002). The Diabetes Prevention Program (DPP) description of lifestyle intervention. *Diabetes care*, 25(12), 2165-2171.
- [16] Lindstrom, J., Louheranta, A., Mannelin, M., Rastas, M., Salminen, V., Eriksson, J., ... & Finnish Diabetes Prevention Study Group. (2003). The Finnish Diabetes Prevention Study (DPS) Lifestyle intervention and 3-year results on diet and physical activity. *Diabetes care*, 26(12), 3230-3236.
- [17] Ramachandran, A., Snehalatha, C., Mary, S., Mukesh, B., Bhaskar, A. D., Vijay, V., & Indian Diabetes Prevention Programme (IDPP). (2006). The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia*, 49(2), 289-297.
- [18] Diabetes Prevention Program Research Group. (2003). Costs associated with the primary prevention of type 2 diabetes mellitus in the diabetes prevention program. *Diabetes care*, 26(1), 36-47.
- [19] Stern, M. P., Williams, K., & Haffner, S. M. (2002). Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test?. *Annals of internal medicine*, 136(8), 575-581.
- [20] Chen, L., Magliano, D. J., Balkau, B., Colagiuri, S., Zimmet, P. Z., Tonkin, A. M., ... & Shaw, J. E. (2010). AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. *Medical Journal of Australia*, 192(4), 197-202.
- [21] Lindstrom, J., & Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes care*, 26(3), 725-731.
- [22] Chang, H. Y., Weiner, J. P., Richards, T. M., Bleich, S. N., & Segal, J. B. (2012). Predicting costs with diabetes complications severity index in claims data. *The American journal of managed care*, 18(4), 213-219.
- [23] Lord, J. M., Flight, I. H., & Norman, R. J. (2003). Metformin in polycystic ovary syndrome: systematic review and meta-analysis. *Bmj*, 327(7421), 951.
- [24] Rathmann, W., Kowall, B., Heier, M., Herder, C., Holle, R., Thorand, B., ... & Meisinger, C. (2010). Prediction models for incident type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. *Diabetic medicine*, 27(10), 1116-1123.
- [25] PW, W. (2007). Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med*, 167, 1068-1074.