Impact of machine learning and feature selection on type 2 diabetes risk prediction

Päivi Riihimaa

Faculty of Medicine, Center for Health and Technology, Digital Health Hub, University of Oulu, Oulu, PL, Finland

Correspondence to: Päivi Riihimaa, PhD (Health Data Specialist). Faculty of Medicine, Center for Health and Technology, Digital Health Hub, University of Oulu, Oulu, FI-90014, PL 5000, Finland. Email: paivi.riihimaa@oulu.fi.

Abstract: This survey summarizes the state of the art for type 2 diabetes mellitus (T2DM) prediction and compares the prediction accuracies obtained by conventional statistical regression and machine learning methods, including deep learning. The impact of feature selection and inclusion of clinical and genomic data on T2DM risk prediction accuracy is also reviewed. The results show that there is a tendency that machine learning algorithms outperform logistic regression in the accuracy of T2DM prediction. Inclusion of clinical data and biomarkers to the core feature set improves accuracy, while incorporating genetic markers in the prediction model is still challenging, due to dimensionality problem and the genetic heterogeneity of T2DM.

Keywords: Keywords: Feature selection; machine learning; precision medicine; risk prediction; type 2 diabetes

Received: 30 January 2020. Accepted: 10 May 2020; Published: 30 June 2020.
doi: 10.21037/jmai-20-4
View this article at: http://dx.doi.org/10.21037/jmai-20-4

Introduction

The ultimate goal of precision medicine is to deliver personalized disease risk assessment together with proactive and preventive care for citizens. Personalization is achieved by analyzing and using various data sources, such as clinical, genomic, environmental and lifestyle data. This task is challenging especially for the global public health challenges, such as cardiovascular diseases and type 2 diabetes mellitus (T2DM), due to the fact that the disease phenotype is a result of complex interaction between genes and environmental factors. This review focuses on the T2DM risk prediction.

T2DM is a common metabolic disorder characterized by insulin resistance and impaired pancreatic beta-cell function (1). There are many environmental, lifestyle and genetic factors known to be important in the development and progression of T2DM. A recent umbrella metareview (2) indicated that at least eleven associations presented convincing evidence for high risk of T2DM. Obesity stands out as a major risk factor, together with sedentary behavior and low adherence to a healthy diet. Association of several environmental factors and biomarkers to T2DM was also supported by convincing evidence. These include high levels of serum uric acid, low level of serum vitamin D, high level of serum alanine aminotransferase (ALT), and exposure to high level of ambient air pollution. Other convincing associations were found in the medical history of the patient (e.g., gestational diabetes, age at menarche) and a few psychosocial factors (2).

Quite many T2DM risk assessment tools have been developed worldwide in order to assess T2DM risk at an individual level, with the purpose of deploying preventive strategies early enough to prevent the progression of the disease from prediabetes stage to diabetes. Because of differences in the genetic and lifestyle characteristics, some scores developed for a certain population may not necessarily perform well in other populations (3). Therefore, a plethora of T2DM risk assessment tools have been created worldwide to get the best tool for each population.

The widely used T2DM risk scores were established during 2000 to 2010 and were based on various regression models (4-9). Recent advancements in the fields of artificial intelligence and machine learning have produced a renaissance of new studies aiming at developing better

T2DM prediction models. This study aims to describe recent development in the T2DM risk prediction with machine learning algorithms and the new features applied in these studies. The specific research questions are: "Does deployment of machine learning algorithms provide more accurate prediction score for T2DM, as compared to the established regression methods?" "What kind of features have been added to T2DM risk prediction models to get a better accuracy?"

Methods

The inclusion criteria for the papers in this review were the following: (I) only published papers in refereed journals since 2003 (not conference papers or chapters in books); (II) dataset size was large enough (arbitrary limit was set to 700 persons); (III) results were validated with either an independent test data set or with 5-fold or 10-fold cross-validation and (IV) prediction performance has been reported with a standard measure, such as area under curve (AUC) or accuracy. AUC is the area under the ROC (receiver operating characteristics) curve, which has been proposed as a single-number measure for evaluating the predictive ability of learning algorithms (10). As all reviewed papers did not publish their results as AUC values, also research papers reporting their prediction performance with accuracy values were included in this review. Accuracy is defined as the number of correct predictions (true positives and true negatives) out of total number of predictions.

For machine learning papers with the commonly used Pima Indian diabetes dataset, only the papers with the highest accuracy or AUC for T2DM prediction were picked to this review from a vast number of research reports. For a compilation for T2DM prediction studies with the Pima Indian dataset, see e.g., Varma & Panda's recent comprehensive survey (11).

Results

A survey of regression models for T2DM prediction

Currently there are numerous logistic regression based T2DM risk scores, which are widely used in clinical settings or as web questionnaires (12). These share a common core set of predicting variables for T2DM [e.g., age, gender, body mass index (BMI) or other measure for obesity, family history of diabetes], but vary in the inclusion of clinical, genetic and lifestyle data (*Table 1*). The prediction performance values as measured by self-reported AUC values range from 0.64 to 0.85. Inclusion of more variables to the risk prediction model does not necessarily result in better AUC value, as the best AUC values (4,7) were achieved with only seven and eight features.

A survey of prediction models using machine learning

The number of machine learning based T2DM risk prediction studies has exploded during the last decade, largely because of the fertile co-operation of data analysts with medical experts. For clarity, *Table 2* does not list all the research studies available, but a representative selection of the main studies with large datasets and the best T2DM prediction accuracies.

Discussion

The reviewed T2DM prediction models based on machine learning showed higher classification accuracies than models based on logistic regression, as measured by the reported accuracy and AUC scores. With logistic regression methods, the AUCs were in the range between 65% and 85%, while machine learning methods gave AUCs between 80.8% and 99.36%—and accuracies in the range of 80.8–99.36%. However, the reported performance values are not directly comparable between the reviewed papers, because the validation strategies and datasets were not the same. Direct comparison for the algorithm's and model's prediction accuracy can be done only if the same dataset and validation dataset is used, as in several papers using Pima Indian dataset for benchmarking the prediction accuracies of different algorithms [e.g., in (23)].

Algorithm selection strategies

In summary, most of the known prediction algorithms have been employed in the reviewed studies. Note that *Table 2* lists only the model which gave the best prediction accuracy for T2DM in each paper. Therefore, it seems that support vector machine (SVM) and random forest tend to be overrepresented among the best performers and give better prediction accuracies than the other tested algorithms. Both are known to be robust to data overfitting, but the resulting predicting model can be difficult to interpret (24)—often regarded as a limitation in medical context.

Deep learning neural networks have not yet been extensively used for T2DM prediction. Recently Ayon

Table 1 Regression model based T2DM risk scores in wide clinical use or as web questionnaires

T2DM risk score	Risk scoring algorithm	Features	AUC	Accuracy	Reference
FINDRISK (Finnish)	Logistic regression	Age, BMI, waist circumference, history of antihypertensive drug treatment, history of high blood glucose, physical activity, daily consumption of fruits, berries, or vegetables	0.85	Not reported	(4)
Indian Diabetes Risk Score (IDRC, Indian)	Multiple logistic regression	Age, abdominal obesity, family history of diabetes and physical activity	0.698	0.613	(5)
Framingham simple clinical risk score (USA)	Logistic regression	Age, gender, Fasting Glucose, BMI, HDL (high-density lipoprotein), triglyceride, blood pressure, parental history of diabetes	0.85	Not reported	(6)
Oman risk score (Oman Arabs)	Logistic regression	Age, family history of diabetes, waist circumference, BMI, current hypertension status	0.83	Not reported	(7)
American Diabetes Association (ADA)	Logistic regression	Age, sex, family history of diabetes, history of hypertension, obesity, and physical activity	0.79	Not reported	(8)
Undiagnosed Diabetes mellitus (UDDM, Indonesian)	Multivariate logistic regression	Age, obesity, central obesity, hypertension, and smoking habit	0.64	Not reported	(9)
CanRisk (Canadian risk)	Logistic regression	Age, BMI, waist circumference, physical activity, fruit/ vegetable consumption, history of high blood pressure, history of high blood glucose, family history of diabetes, sex, ethnicity, maternal history of macrosomia, and education	0.75	Not reported	(13)
Register data from an insurance company, Pennsylvania, USA	Logistic regression	900 variables	0.80	Not reported	(14)
T2DM risk score for Chinese rural population (China)	Cox regression	Age, BMI, triglycerides, fasting plasma glucose	0.768	Not reported	(15)

& Islam (21) published promising results in their paper describing the application of deep learning neural network for predicting T2DM in the Pima Indian dataset, with high accuracy (98.0 %, *Table 2*). As Pima Indians have limited genetic and environmental variability (25), the prediction task is far easier than in other, genetically and environmentally more complex populations. It will be interesting to see how deep learning approach performs in future studies with larger datasets of mixed ethnicity.

Deep learning is widely used in the analysis of complex medical data, especially in medical imaging, as they have the benefit of using nonlinear activation functions instead of linear ones. In fact, in clinical health datasets the relationships between the dependent and independent variables are commonly not linear, such as gene-environment interactions or association between various physiological features, such as obesity and hypertension.

Feature selection

In addition to proper selection of the algorithm, accuracy of the disease prediction model can be improved by (I) increasing dataset size to capture enough variation of the population and (II) by careful selection of the most relevant features to the prediction model.

The core features for T2DM prediction were shared by majority of the reviewed papers, including age, BMI, waist circumference, hypertension and family history of diabetes. These features can be easily gathered in web surveys without any clinical data, thereby forming a general basis for T2DM prediction tools. Addition of clinical data (triglycerides, fasting plasma glucose, cholesterol data etc.) improved T2DM risk score performance (as measured by AUC) up to 85% for logistic regression models and up to 98% for machine learning based models.

Table 2 Machine learning based T2DM risk scores

Dataset	Size (number)	Features	Algorithm	Validation strategy	AUC	Accuracy	Reference
Pima Indian dataset	768	Age, BMI, number of pregnancies, glucose tolerance test result, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, diabetes pedigree function	Neuro fuzzy inference	10-fold cross-validation	92.0	89.47	(16)
NHANES (National Health and Nutrition Examination Survey) USA	6,314	Family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, physical activity, smoking, alcohol use, education, and household income	Support Vector Machine	10-fold cross-validation	83.5	Not reported	(17)
Hospital in India	1,415	Age, BMI, waist circumference, diastolic blood pressure, hypertension, lifestyle, gender, high cholesterol, physical exercise, smoking, food type (vegetarian/nonvegetarian), major cereal (rice/wheat), family history of diabetes, drinking habit, class	Several combined to an expert model	Separate test dataset partitioned from the original dataset (311 out of 1,415)	Not reported	99.36	(18)
Kuwait Health Network	10,632	height, weight, age, gender, ethnicity, hypertension diagnosis and a family history of hypertension and diabetes	Support Vector Machine	5-fold cross-validation	Not reported	81.3 in general population, higher in specific subpopulations	(3)
Girona, Spain	828	Genetic data (Single nucleotide polymorphisms, SNP), age, BMI, sex	Random forest	10-fold cross-validation	89.0	Not reported	(19)
Chinese hospital	68,994	Age, pulse rate, breathe, left systolic pressure (LSP), right systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, physique index, fasting glucose, waistline, low-density lipoprotein (LDL), and high-density lipoprotein (HDL)	Random forest	5-fold cross-validation	Not reported	80.8	(20)
Pima Indian dataset	768	Age, BMI, number of pregnancies, glucose tolerance test result, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, diabetes pedigree function	Deep learning neural network	5-fold cross-validation	98.0	98.0	(21)
NHANES (National Health and Nutrition Examination Survey, USA)	>5,000	24 most important variables out of 123: blood osmolality, sodium, blood urea nitrogen, triglyceride, LDL, age, waist, leg length, chloride, self-reported greatest weight, close relative had diabetes, total cholesterol, gamma glutamyl transferase, ethnicity, systolic blood pressure, HDL, pulse, carbohydrate intake, general health condition, mean cell volume, aspartate aminotransferase (AST), lymphocyte number, white blood cell count	XGBoost (an ENSEMBL model)	10-fold cross-validation	84.4	Not reported	(22)

Performance values are from the best model in case several models were presented in the paper. AST, aspartate aminotransferase; AUC, area under the curve; BMI, body mass index; BXGBoost, extreme gradient boost; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LSP, left systolic pressure; LDP, left diastolic pressure; RDP, right diastolic pressure; RSP, right systolic pressure.

Impact of genetic data

Genetic data can be included in the feature set for T2DM risk prediction either as a simple variable describing the family history of diabetes or as precise genetic markers associated with the disease, such as small nucleotide polymorphisms (SNPs). So far, SNPs have been included in the feature set for T2DM risk prediction in only few papers using machine learning approach. Lopez *et al.* (19) achieved an intermediate AUC of 0.85 for T2DM prediction by including SNP data to their predictive model together with only a few basic variables, but without clinical or biomarker data. The challenge with SNP data is a dimensionality problem: hundreds of subjects with thousands of SNPs per subject, resulting in complex models and high demands for computing capacity. Another challenge with SNPs is model overfitting (19), resulting in suboptimal generalizability of the prediction model to other datasets.

To date, over 400 genetic variants have been associated with the onset of T2DM, but even combined, they explain just 18% of the risk of T2DM, and the risk associated with individual variants is low, usually less than 1.2% (25,26). In a study with only genetic (SNP) data for T2DM risk prediction, different SNP combinations achieved prediction accuracy of 70.9% with Support Vector Machine algorithm for men in Korean subpopulations (27), emphasizing the inclusion of genetic markers to the disease prediction models.

Genetic heterogeneity within T2DM complicates the situation for T2DM risk prediction based on genetic data. A data-driven cluster analysis (k-means and hierarchical clustering) in patients with newly diagnosed diabetes in a Swedish cohort identified five replicable clusters of patients with diabetes, which had significantly different patient characteristics and risk of diabetic complications (28). When integrating SNP variant data to the T2DM risk prediction model, it might be interesting to first classify the different subgroups and look for the predicting variables and the best predicting model separately for each of these.

Acknowledgment

The authors thank the European Regional Development Fund, University of Oulu, Faculty of Medicine and the City of Oulu for their support in this study.
Funding: None.

Footnote

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai-20-4). The author has no conflicts of interest to declare.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Petersen KF, Dufour S, Morino K, et al. Reversal of muscle insulin resistance by weight reduction in young, lean, insulin-resistant offspring of parents with type 2 diabetes. Proc Natl Acad Sci USA 2012;109:8236-40.
2. Bellou V, Belbasis L, Tzoulaki I, et al. Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of metaanalyses. PLoS One 2018;13:e0194127.
3. Farran B, Channanath AM, Behbehani K, et al. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. BMJ Open 2013;3:e002457.
4. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003;26:725-31.
5. Mohan V, Deepa R, Deepa M, et al. A simplified Indian diabetes risk score for screening for undiagnosed diabetic subjects. J Assoc Physicians India 2005;53:759-63.
6. Wilson PW, Meigs JB, Sullivan L, et al. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 2007;167:1068-74.

7. Al-Lawati JA, Tuomilehto J. Diabetes risk score in Oman: a tool to identify prevalent type 2 diabetes among Arabs of the Middle East. Diabetes Res Clin Pract 2007;77:438-44.
8. Bang H, Edwards AM, Bomback AS, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009;151:775-83.
9. Pramono LA, Setiati S, Soewondo P, et al. Prevalence and predictors of undiagnosed diabetes mellitus in Indonesia. Acta Med Indones 2010;42:216-23.
10. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299-310.
11. Varma M, Panda M. Comparative analysis of Predicting Diabetes Using Machine Learning Techniques. J Emerg Technol Innov Res 2019;6:522-30.
12. Buijsse B, Simmons RK, Griffin SJ, et al. Risk Assessment Tools for Identifying Individuals at Risk of Developing Type 2 Diabetes. Epidemiol Rev 2011;33:46-62.
13. Robinson CA, Agarwal G, Nerenberg K. Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. Chronic Dis Inj Can 2011;32:19-31.
14. Razavian N, Blecker S, Schmidt AM, et al. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data 2015;3:277-87.
15. Zhang M, Zhang H, Wang C, et al. Development and Validation of a Risk-Score Model for Type 2 Diabetes: A Cohort Study of a Rural Adult Chinese Population. PLoS One 2016;11:e0152054.
16. Polat K, Günes S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digit Signal Process 2007;17:702-10.
17. Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010;10:16.
18. Shankaracharya, Odedra D, Samanta S, et al. Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. Rev Diabet Stud 2012;9:55-62.
19. López B, Torrent-Fontbona F, Viñas R, et al. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif Intell Med 2018;85:43-49.
20. Zou Q, Qu K, Luo Y, et al. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet 2018;9:515.
21. Ayon SI, Islam MM. Diabetes prediction: a deep learning approach. Int J Inform Engr Electr Business 2019;2:21-7.
22. Dinh A, Miertschin S, Young A, et al. A data-driven approach for predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 2019;19:211-26.
23. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Computing Informatics 2018. doi: 10.1016/j.aci.2018.12.004.
24. Dankwa-Mullan I, Rivo M, Sepulveda M, et al. Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here. Popul Health Manag 2019;22:229-42.
25. Mahajan A, Wessel J, Willems SM et al. Refining the accuracy of validated target identification through coding variant finemapping in type 2 diabetes. Nat Genet 2018; 50:559-71.
26. Prasad RB, Groop L. Precision medicine in type 2 diabetes. J Internal Med 2019;285:40-8.
27. Ban HJ, Heo JY, Oh KS, et al. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC Genet 2010;11:26-37.
28. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 2018;6:361-9.

doi: 10.21037/jmai-20-4

Cite this article as: Riihimaa P. Impact of machine learning and feature selection on type 2 diabetes risk prediction. J Med Artif Intell 2020;3:10.