diabetology

MDPI

Commentary

An Online Risk Tool for Predicting Type 2 Diabetes Mellitus

Gian Alix [1], Huaxiong Huang [2,3,4,5], Aziz Guergachi [6], Karim Keshavjee [7] and Xin Gao [4,*]

1 Department of Electrical Engineering & Computer Science, Lassonde School of Engineering, York University, Toronto, ON M3J1P3, Canada; galix@yorku.ca
2 Research Center of Mathematics, Advanced Institute of Natural Sciences, Beijing Normal University at Zhuhai 519087, China; hhuang@yorku.ca
3 BNU-HKBU United International College, Zhuhai 519087, China
4 Department of Mathematics & Statistics, York University, Toronto, M3J1P3 ON, Canada
5 Fields CQAM Health Analytics and Multidisciplinary Modeling Lab, Toronto, M5T3J1 ON, Canada
6 Ted Rogers School of Management, Ryerson University, Toronto, ON M5B2K3, Canada; a2guerga@ryerson.ca
7 Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON M5T3M6, Canada; karim.keshavjee@ryerson.ca
* Correspondence: xingao@yorku.ca

Abstract: An online risk prediction tool is developed to calculate a user's risk of developing type II diabetes mellitus (T2DM). The risk prediction is based on the user's input of medical lab information, such as age, sex, body mass index, fasting blood sugar, triglycerides, and high-density lipoprotein levels. The calculator is modelled using a logistic regression model, and it is trained using the medical records of over ten thousand Canadian patients. This newly developed tool is intended to serve physicians and patients in predicting future diabetes risk and take early preventive measures.

Keywords: diabetes mellitus; logistic regression model; machine learning; predictive models; online risk tool

Citation: Alix, G.; Huang, H.; Guergachi, A.; Keshavjee, K.; Gao, X. An Online Risk Tool for Predicting Type 2 Diabetes Mellitus. *Diabetology* **2021**, *2*, 123–129. https://doi.org/10.3390/diabetology2030011

Academic Editor: Peter Clifton

Received: 19 February 2021
Accepted: 22 June 2021
Published: 8 July 2021

1. Development of the Online Risk Tool

Diabetes has continuously proven to be a challenging chronic disease for the general population. In the machine learning era, predictive models have been proposed in the literature to predict diabetes risk based on various types of predictors. For instance, Tigga and Garg [1] assessed the performance of a random forest classifier to predict type 2 diabetes in the Indian community, based on health, lifestyle and family background factors. The model was applied to the Pima Indian Diabetes data and showed results with a specificity of 66.1% and a sensitivity of 78.9%. Kopitar et al. [2] investigated the early detection of T2DM using machine learning-based prediction models given 6 months of available data. Zhang et al. [3] used machine learning models such as logistic regression, random forest, and gradient boosting machines to predict type 2 diabetes in a rural Chinese population. De Silva et al. [4] conducted a systematic review and meta-analysis of more than 20 machine learning models that are tasked with predicting the risk of type 2 diabetes in a community setting. The primary objective of the research study was to determine the predictive power and performance of the identified machine learning models. The results show that the models performed well in terms of predicting the presence of diabetes in patients. However, there is a lack of online tools which implement the existing machine learning models for practitioner use. In this paper, we introduce an online risk prediction tool for type 2 diabetes developed by a team of experts with different backgrounds, such as doctors, clinical IT architects, statisticians and computer scientists. It is hoped that the new tool will bring forth significant contributions to early disease prevention and personalized medical care for diabetes patients.

Hang et al. [5] proposed effective predictive models including a logistic regression model and a gradient boosting machine model, to predict the risk of type 2 diabetes

mellitus in Canadian patients. To make the predictive tool accessible to physicians and patients, we developed a web application that implements the logistic regression model in [5]. This is an online risk tool for predicting type 2 diabetes mellitus, where a user may input his/her laboratory information to obtain the risk prediction (http://www.yorku.ca/xingao/t2diabetesPredictor.html, accessed on 29 June 2021). The model is trained from the electronic records of Canadian patients obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). The dataset includes patients ranging from 18 to 90 years old, along with their laboratory information which are used as the predictors of the disease: age (at time of examination), sex, fasting blood glucose, body mass index, lipoprotein levels (high- and low-density), triglycerides, and systolic blood pressure. Figure 1 outlines the criteria for how these predictors were selected via a flowchart. The dataset used by the predictive model was curated through a process that is summarized in Figure 2. For healthy individuals, we used the last visit in their record; for diabetes patients, we used the last visit before their disease onset time. In the end, 13,309 records were used by the predictive model.

Approximately 20.9% of the records exhibit disease status with type 2 diabetes mellitus. In the data, 40% of the observations are male while 60% are female. Ages were categorized according to the following groups: young (<40 years old), middle-aged (40–64 years old), senior (65–84 years old), and elderly (>85 years old). About 44.6% of the patients make up the middle-age category, 47.8% are seniors, 4.8% are elderly, and 2.9% belong to the young group. The body mass index is calculated as the result of dividing a patient's weight (in kilograms) by the square of the height (in meters). It was observed that the BMI ranges from 11.2 to 70 among patients, with a median of 28.9. The distributions of BMI, fasting blood glucose, high-density lipoprotein and triglycerides are right-skewed. The most important predictors in predicting Type II diabetes risk are fasting blood glucose, body mass index, high-density lipoprotein and triglycerides, each factor having a p value < 0.0001. In fact, the results show that these predictors were all strongly linearly associated with the outcome of diabetes mellitus on the logit scale. The coefficients of these variables were estimated to be 1.963, 0.023, −0.894, and 0.158, respectively, while the odds ratio of these predictors were calculated as 7.122, 1.024, 0.409, and 1.171, respectively. It was also shown that age proved to be a significant factor as well. With the middle-aged group as the reference, the elderly group has a p value of <0.0001, the senior group with a p value of 0.036 and the young group with a p value of 0.170. These three age groups have odds ratio of 0.436, 0.881, and 1.269 compared to the middle-aged group. Sex is also shown to be a contributing factor that influences the disease risk with a p value < 0.0001; males have an estimated coefficient of −0.250 with odds ratio 0.779 compared to female group. Meanwhile, the predictors systolic blood pressure and low-density lipoprotein proved to be insignificant to the model. We also provide the 95% confidence intervals for the odds ratio and the average predicted probability for each predictor in Table 1. In checking for the assumptions of the model, no severe collinearities were found and that the computed variance inflation factor (VIF) was shown to be less than 1.5.

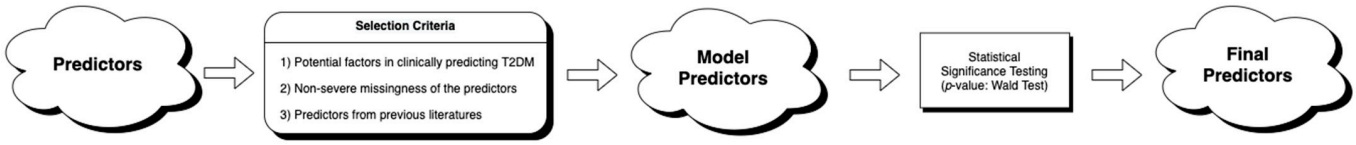

Figure 1. Flowchart exhibiting the process of how the predictors were selected. From the predictors of the original dataset provided by the CPCSSN, 8 predictors were selected based on the three criteria listed. These chosen predictors include: age, sex, fasting blood glucose, body mass index, (high- and low-density) lipoprotein levels, triglycerides, and systolic blood pressure. However, two of them (low-density lipoprotein and systolic blood pressure) were deemed insignificant based on statistical testing (via the p-value based on the Wald Test). The remaining predictors were then considered to be the final predictors of the model.

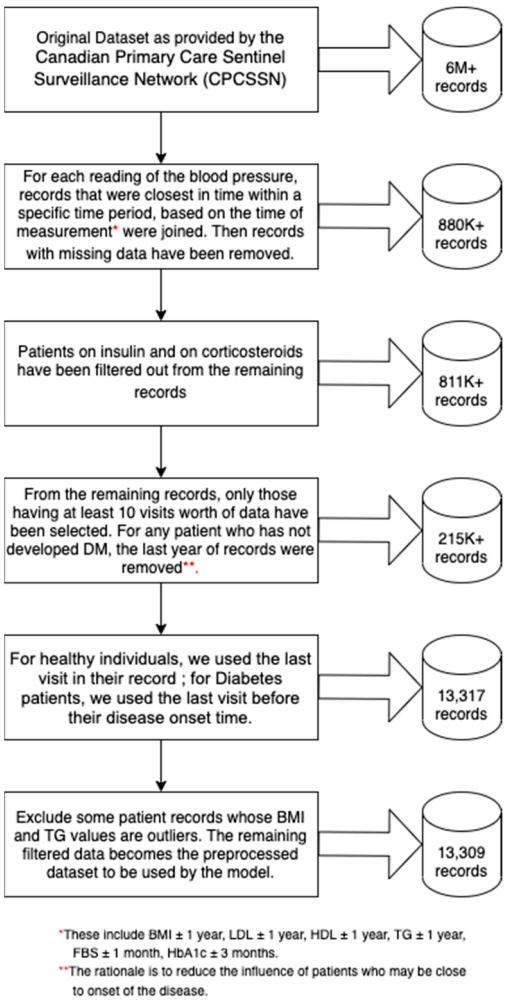

Figure 2. Flowchart exhibiting the process of how the dataset was curated. At each step of the filtering process, the number of records remaining was noted. From 6 million records in the original dataset, 13,309 remained to be used by the predictive model.

Table 1. The 95% confidence intervals of odds ratio and the average predicted probability for each predictor.

Predictor	95% Confidence Interval	Average Predicted Probability
Age		
Elderly	(0.3470, 0.6639)	−0.0886
Senior	(0.8367, 1.0308)	−0.0090
Young	(0.8409, 1.5036)	0.0149
Male	(0.7062, 0.8748)	−0.0294
FBS	(6.2638, 7.4771)	0.2347
BMI	(1.0208, 1.0365)	0.0034
HDL	(0.3760, 0.5231)	−0.0991
TG	(1.0696, 1.2107)	0.0158
sBP	(0.9937, 1.000)	−0.0004
LDL	(0.9168, 1.0227)	−0.0039

Our online risk tool was built using HTML5 as the main framework of the web application, front end CSS3 for the site's style design, and back end JavaScript for the application's functionality. See Appendix A for a more thorough discussion on the web development of the application. The prediction tool, through a simple form, prompts the needed information from the user; the results can then be displayed and interpreted immediately. Our online web tool also provides brief descriptions of each parameter being prompted. Validated on the data set of 13,309 Canadian patients, our web tool achieves satisfactory results. The prediction tool has a sensitivity of 77.3%, a specificity of 71.3%, and an area under the receiver-operating characteristic curve of 0.74, based on a 0.20 threshold value. Steyerberg et al. [6] specified other specific metrics including balanced accuracy, calibration slope and discrimination slope to evaluate a classification method. For our model, the balanced accuracy is 0.743, the calibration slope is 1.01 and the discrimination slop is 0.26, respectively. The plots for the calibration line and the discrimination box are provided in Figure 3.

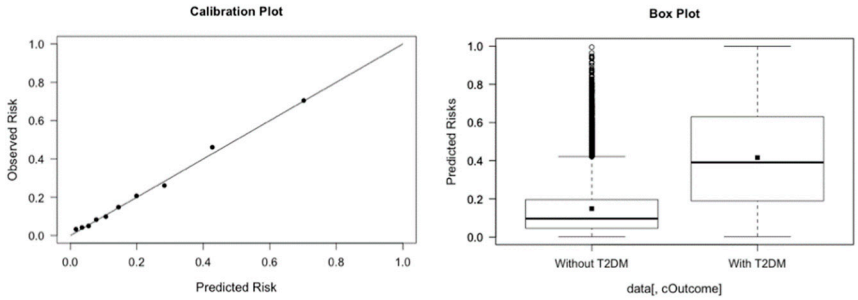

Figure 3. The calibration plot and the discrimination box plot of the logistic model.

To use the online risk tool, the user enters the predictor values which are readily available from the patient's lab results. The online tool predicts and displays the future risk of type 2 diabetes mellitus for the user. This tool can display the user's risk in terms of risk likelihood (0–1), risk level: insignificant risk of DM (likelihood < 0.01), low risk of DM (0.01 < likelihood < 0.20); moderate risk of DM (0.20 < likelihood < 0.50); moderate-high risk of DM (0.50 < likelihood < 0.75); high risk of DM (likelihood > 0.75)), and an associated message based on the risk level. To determine the cutoffs of likelihood for different risk levels, we considered the 0.20 cutoff value which was be determined by the best combination of sensitivity and specificity metrics. Then, we used this 0.20 to dichotomize between the "no risk" and the "risk" group. Within the group having no risk, we further subdivided the groups into "insignificant risk" and "low risk". We used 0.01 as the cutoff. Similarly, in the group with risk, we used 0.50 and 0.75 as the cutoff values and further subdivided into three groups. We use the finer partition of five risk categories to provide more information on the level of the risk.

We compute the odds of developing the disease for the five groups as follows:

$$\text{odds} = \frac{\Pr(\text{ with T2DM} \mid \text{risk category})}{\Pr(\text{ without T2DM} \mid \text{risk category})}$$

where the risk categories are the five aforementioned risk groups. The values of the odds are 0.0417, 0.0989, 0.4706, 1.7027, and 4.8824, for insignificant risk, low risk, moderate risk, moderate-high risk, and high risk of DM, as summarized in Table 2. It can be observed that the odds of developing the disease for the insignificant and low risk groups are rather low, moderate for the moderate risk group and especially elevated for the moderate-high and high risk groups.

Table 2. Percentage breakdown of patients with/without the disease among the 5 risk categories. The third row provides the odds of getting the T2DM disease given the risk of each category.

	Insignificant Risk	Low Risk	Moderate Risk	Moderate-High Risk	High Risk
With T2DM	3.659%	8.702%	32.21%	63.37%	82.77%
Without T2DM	96.34%	91.30%	67.79%	36.63%	17.23%
Odds	0.0417	0.0989	0.4706	1.7027	4.8824

Currently, several diabetes risk calculators are available online. While our calculator is clinical and is based on blood work, the Finnish diabetes risk calculator FINDRISC by QxMD uses various genetic and dietary-based predictors such as the patients' gender, age, dietary and exercise habits, parental disease status, weight and height as predictors. This calculator performed at 72% specificity and at 77% sensitivity [7]. The Canadian Public Health Agency has an online diabetes risk calculator CanRISK [8], which is similar to FINDRISC and uses gender, age, family history, ethno-cultural background, weight, height, waist circumference, physical activity, diet, blood pressure, and blood sugar level as risk factors. The Diabetes Risk Calculator [9] developed by the Omni Calculator Project takes into account ethnicity and family history factors in addition to the typical factors (i.e., sex, age, blood pressure, cholesterol, etc.). Other existing online risk calculators that were developed as an initiative to reduce type 2 diabetes risk include the Australian Type 2 Diabetes Risk Assessment Tool [10] by the Baker IDI Heart Diabetes Institute and the 60-Second Type 2 Diabetes Risk Test [11] by the American Diabetes Association. What distinguishes our prediction tool from the other risk prediction calculators is the fact that this tool is modelled using routine laboratory results and the model is specifically trained using records of Canadian patient laboratory results ordered by their family doctors. The tool is trained and validated on a Canadian population. Figure 4 displays the graphical interface of the online risk prediction tool. In our future work, we intend to consider any other predictors as well, such as exercise habits, family history, and ethnic background to further improve the performance of our risk tool. Another limitation of our online risk calculator is the fact that in the data curation process, some patients with borderline diabetes are removed. As a future work, we plan to include more borderline patients so that the re-trained model could be used to identify borderline disease patients.

Currently, the data used for training and validating the model by our risk tool are not stored online. Another future work that we intend to complete on our calculator is to store all of these data, including any additional data that we collect in future into a working real-time database. In this manner, we aim to construct our model to continuously "learn" and produce progressively accurate and more improved results overtime. Additionally, we aim to investigate how to use a risk tool to predict individual disease on-set time.

Several research works in the literature have assessed the performances of various machine learning models in predicting the risk of a person developing type 2 diabetes mellitus. However, the risk calculators that physicians can easily use are still inadequate. Hence, we develop an online risk calculator to predict the risk of T2DM based on a user's medical lab data (age, sex, BMI, fasting blood sugar levels, triglycerides, and high-density lipoprotein levels). Our tool is modelled using routine laboratory results and is trained and validated on a Canadian population. With this online tool, we intend to support physicians and patients in predicting the incidence of diabetes, thus allowing our medical experts to provide timely preventive and personalized interventions.

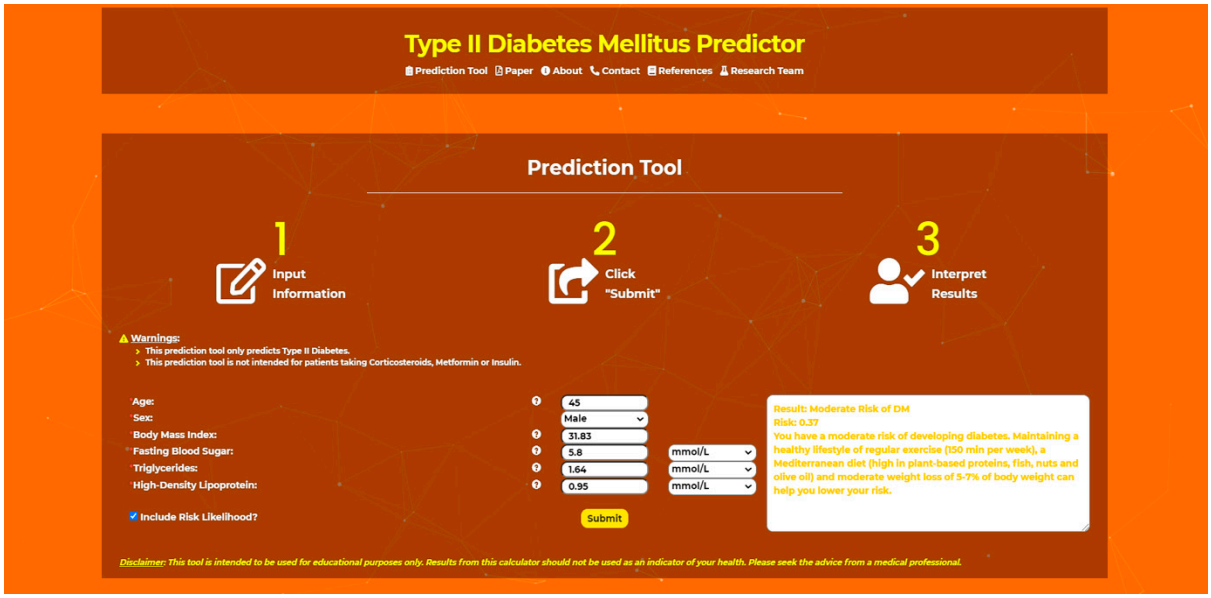

Figure 4. Graphical interface of the online risk prediction tool for type 2 diabetes mellitus.

Author Contributions: Conceptualization, X.G., H.H., A.G. and K.K.; methodology, G.A. and X.G.; software, G.A.; validation, G.A.; formal analysis, G.A.; investigation, G.A.; resources, X.G. and H.H.; data curation, A.G. and K.K.; writing—original draft preparation, G.A.; writing—review and editing, X.G.; visualization, G.A.; supervision, X.G.; project administration, X.G.; funding acquisition, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research work of X.G. is supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

Data Availability Statement: The data that support the findings of this study are available from CPCSSN (www.cpcssn.ca, accessed on 29 June 2021) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of CPCSSN.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Web Development Process and Design

We outline how our online risk tool was developed. The main framework was built from scratch with hypertext markup language (HTML5). Cascading style sheets (CSS3)-based simplistic styles were applied to modify the site's fonts, colors, and backgrounds. The functionality of the risk tool was built with the easy-to-program Javascript framework. Both CSS3 and Javascript were jointly used in the production of node trajectories in the background, as part of the website's design. Then, to enable our risk tool to be smooth, seamless, and responsive, particularly on mobile browsers, we employed a Bootstrap3 framework. This framework supports a grid system that allows the website to be flawless and structured, regardless of whether viewed under a desktop web browser or a mobile web browser—a feature not possible without Boostrap3. Glyphicons (or icons) from Bootstrap3 have also been used, in addition to some of the ones imported from FontAwesome (https://fontawesome.com/icons?d=gallery&p=2, accessed on 29 June 2021). We foresee the continuous development of our online tool, based on our intended future work.

References

1. Tigga, N.P.; Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Comput. Sci.* **2020**, *167*, 706. [CrossRef]
2. Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; Sitglic, C. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **2020**, *10*, 11981. [CrossRef] [PubMed]

3. Zhang, L.; Wang, Y.; Niu, M.; Wang, C.; Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Sci. Rep.* **2020**, *10*, 4406. [CrossRef]
4. De Silva, K.; Lee, W.K.; Forbes, A.; Denmer, R.; Barton, C.; Enticott, J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *IJMI* **2020**, *143*, 104268. [CrossRef]
5. Lai, H.; Huang, H.; Keshavjee, K.; Guergachi, A.; Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord.* **2019**, *19*, 101. [CrossRef] [PubMed]
6. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–138. [CrossRef] [PubMed]
7. Calculate by QxMD, FINDRISC Diabetes Risk Calculator. Available online: https://qxmd.com/calculate/calculator_675/diabetes-risk-score-type-2 (accessed on 29 June 2021).
8. Canadian Public Healthy Agency. The Canadian Diabetes Risk Questionnaire. Available online: https://canadiantaskforce.ca/tools-resources/type-2-diabetes-2/type-2-diabetes-canrisk/ (accessed on 29 June 2021).
9. Omni Calculator. Diabetes Risk Calculator. Available online: https://www.omnicalculator.com/health/risk-dm (accessed on 29 June 2021).
10. Diabetes Australia. Diabetes Risk Calculator. Available online: https://www.diabetesaustralia.com.au/risk-calculator/ (accessed on 29 June 2021).
11. American Diabetes Association. Our 60-Second Type 2 Diabetes Risk Test. Available online: https://www.diabetes.org/risk-test (accessed on 29 June 2021).