Machine Learning Algorithms for Prediabetes Risk Calculation: a Protocol of Systematic Review and Meta-Analysis

Yaltafit Abror Jeem (yaltafit.abror.j@ugm.ac.id)
Islamic University of Indonesia https://orcid.org/0000-0003-1388-2922
Refa Nabila
Universitas Islam Indonesia Integrated Campus: Universitas Islam Indonesia
Dwi Ditha Emelia
Universitas Islam Indonesia Integrated Campus: Universitas Islam Indonesia
Lutfan Lazuardi
Gadjah Mada University Faculty of Medicine: Universitas Gadjah Mada Fakultas Kedokteran Kesehatan Masyarakat dan Keperawatan
Hari Kusnanto Josef
Gadjah Mada University Faculty of Medicine: Universitas Gadjah Mada Fakultas Kedokteran Kesehatan Masyarakat dan Keperawatan

Methodology

Keywords: Prediabetic state, Telemedicine, Machine Learning, Screening, Diagnostic, Systematic Review

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-578915/v1

Abstract

Background One strategy to resolve the increasing prevalence of T2DM is to identify and administer interventions to prediabetes patients. Risk assessment tools help detect diseases, by allowing screening to the high risk group. Machine learning is also used to help diagnosis and identification of prediabetes. This review aims to determine the diagnostic test accuracy of various machine learning algorithms for calculating prediabetes risk.

Methods This protocol was written in compliance with the Preferred Reporting Items for Systematic Review and Meta-Analysis for Protocols (PRISMA-P) statement. The databases that will be used include PubMed, ProQuest and EBSCO restricted to January 1999 and May 2019 in English language only. Identification of articles will be done independently by two reviewers through the titles, the abstracts, and then the full-text-articles. Any disagreement will be resolved by consensus. The Newcastle-Ottawa Quality Assessment Scale will be used to measure the quality and potential of bias. Data extraction and content analysis will be performed systematically. Quantitative data will be visualized using a forest plot with the 95% Confidence Intervals. The diagnostic test outcome will be described by the summary receiver operating characteristic curve. Data will be analyzed using Review Manager 5.3 (RevMan 5.3) software package.

Discussion We will obtain diagnostic accuracy of various machine learning algorithms for prediabetes risk estimation using this proposed systematic review and meta-analysis.

Systematic review registration: This protocol has been registered in the Prospective Registry of Systematic Review (PROSPERO) database. The registration number is CRD42021251242.

Background

The type 2 diabetes mellitus (T2DM) is becoming more common while the human, financial and social situations are deteriorating (1). The number of T2DM cases, projected to increase to over 693 million by 2045, represent one of the main public health issues and impacted nearly 451 million individuals globally in 2017 (2). T2DM is linked to several long-term complications including heart failure, visual impairment, and kidney disease (3).

Prediabetes is a significant factor for the progression of T2DM. Most individuals go through a prediabetes phase before full-blown T2DM develops (4, 5). Impaired fasting glucose (IFG), impaired glucose tolerance (IGT), and/or elevated HbA1c levels between 5.7% and 6.4% are used to diagnose prediabetes (6). It has been shown that people with prediabetes are more likely to develop T2DM than people with normal blood glucose levels. In comparison to normoglycemic people, those with isolated IGT have more than five times the risk, those with isolated IFG have seven times the risk, and those with IGT and IFG have more than twelve times the risk of developing T2DM within a year (7).

Importantly, one strategy to resolve the high burden of T2DM is to detect and offer early interventions to those with prediabetes. The possibility that a certain health outcome will be expected in view of the person's characteristics (risk factors) may be developed in a risk assessment tool. In screening to highest risk group, risk assessment tools help to maximize the available resources for detecting a disease (8). As a result, risk evaluation are useful and ethical in identifying those with prediabetes, who may benefit from interventions, with many health researchers recommending them as the first step in a screening program (9).

The cause of prediabetes is multifactorial. Several risk factors for prediabetes have been known nowadays. The risk factors for prediabetes are older age, sex, history of gestational DM or having given birth to a baby weighing more than 9 pounds., family history of DM (sibling's DM and parent's DM), hypertension, physical inactivity, high triglyceride level, abdominal obesity and obesity (10, 11). The algorithm or early detection model of determinant factors can be used to calculate prediabetes risk factors. The algorithm calculation is used to diagnose the risk of prediabetes and is a prevention action, using machine learning to classify the predictor variables of the prediabetes risk. Machine learning classification is a form of artificial intelligence (AI) that allows computers to learn without specifically being programmed. It has been used to develop a scoring method for prediabetes identification and diagnosis (12, 13).

Machine learning is aimed at developing algorithms that enable computers to study and gain mastery or applicable understanding depending on their previous experiences. This branch is strongly associated with statistics and is a branch of AI. By mastering, it implies that the model can recognize and understand the information, so that decisions and predictions can be developed on this basis (14). Machine learning is usually divided into three main classifications, including: supervised learning, unsupervised learning, and reinforcement learning. Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Genetic Algorithms (GA), and k-Nearest Neighbors (k-NN), which are some of the most popular techniques (15). This review aims to determine the diagnostic test accuracy of various machine learning algorithms for calculating prediabetes risk.

Materials And Methods

Study design and research variable

This protocol was written in compliance with the requirements of the Preferred Reporting Items for Systematic Review and Meta-Analysis for Protocols (PRISMA-P) statement. (Additional file 1) (16). This proposed protocol for review of machine learning algorithms for prediabetes risk calculation has been registered in the Prospective Registry of Systematic Review (PROSPERO) database. The registration number is CRD42021251242. This study will be done as a quantitative study with a meta-analysis study design, if applicable. Meta-analysis will be used to figure the diagnostic accuracy in numerous machine learning algorithms.

The variables of this study include the independent variables consisting of Machine learning algorithms such as Genetic Algorithms, Naive Bayes, Super Vector Machine, Neural Network, Decision Tree, Linear Discriminant Analysis, Logistic Regression, and K-Nearest Neighbor; and the dependent variable is prediabetes.

Research Procedure

This study will be conducted by gathering information and identifying published research papers in the in PubMed, ProQuest and EBSCO article databases on diagnostic test precision of different machine learning algorithms for prediabetes risk calculation.

Study selection will be done by online searching using following keywords: ((prediabetes risk OR prediabetes risk calculation OR prediabetes prediction) AND (machine learning OR algorithms OR Naive Bayes OR Neural Network OR Decision Tree OR Logistic Regression OR Linear Discriminant Analysis OR Super Vector Machine OR K-Nearest Neighbor OR Genetic Algorithms)).

The search will be restricted only for articles in the English language. The topic of the study will be restricted to human subject studies. Publication time will be restricted from January 1999 to May 2019. The inclusion criteria for this study sample are research on machine learning algorithms for prediabetes risk calculations with prognostic model study. The articles will be omitted if they are: (a) irrelevant, (b) not studying machine learning algorithm, (c) the provided information was inadequate and (d) duplicate studies.

Study selection

Initially, the search results will be submitted to the citation management database. This reference manager will identify and delete duplicates through the entire search. Two reviewers (RN and DDE) will identify papers by reviewing the titles, then the abstracts and finally the full-text forms. Potentially relevant studies in full-text articles will be retrieved by all authors. In case of a disagreement, the decision will be reached by consensus.

Data collection technique

Data from the included studies will be compiled into an Excel table spreadsheet by two reviewers (RN and DDE) independently. From each article, data such as first author's name and year of publishing, dataset size, dataset, machine learning algorithms and outcome will be gathered. Area Under Curve (AUC), Sensitivity (Se), specificity (Sp), Positive Predictive Value (PPV), and Negative Predictive Value (NPV) values from each article will be recorded as the main outcomes. If the desired data are not reported in an article, we will contact the corresponding author of the article to retrieve any missing data. Additional file 3 provides the data extraction form.

Quality assessment

Quality assessment of the article will be performed using Newcastle–Ottawa Quality Assessment Scale (NOS). Articles will be assessed as high quality if the NOS score ≥ 7 (17). The results of quality assessment will be presented in summary tables with other characteristic information from the articles.

Data analysis

Raw data will be used to produce the paired forest plot. False negative (FN), false positive (FP), true negative (TN) and true positive (TP) will be used to calculate sensitivity (Se) and Specificity (Sp). The diagnostic test outcome will be described by the summary receiver operating characteristic (SROC) curve. AUC results serve as a rough guide for classifying the diagnostic test accuracy. The criteria for AUC classification are 0.90-1 (excellence), 0.80–0.90 (good), 0.70–0.80 (fair), 0.60–0.70 (poor) and 0.50–0.60 (failure) (18). The Review Manager 5.3 (RevMan 5.3) software package will be used to evaluate the data.

Discussion

Machine learning has mainly been used to help the diagnosis and identification of prediabetes. We will obtain diagnostic accuracy of various machine learning algorithms for prediabetes risk estimation using this proposed systematic review and meta-analysis.

List Of Abbreviations

AI: Artificial Intelligence; ANN: Artificial Neural Networks; AUC: Area Under Curve; CIs: confidence intervals; DM: Diabetes Mellitus; DT: Decision Trees; FN: False Negative; FP; False Positive; GA: Genetic Algorithms; HbA1c: Hemoglobin A1c; IBL: Instance Based Learning; IFG: Impaired Fasting Glucose; IGT: Impaired Glucose Tolerance; k-NN: k-Nearest Neighbors; NOS: Newcastle–Ottawa Quality Assessment Scale; NPV: Negative Predictive Value; PPV: Positive Predictive Value; PRISMA-P: Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocol; PROSPERO: Prospective Registry of Systematic Review; Se: Sensitivity; Sp: Specificity; SROC: Summary Receiver Operating Characteristic; SVM: Support Vector Machines; TD2M: Type-2 Diabetes Mellitus; TN: True Negative; TP: True Positive.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YAJ was responsible for the formulation of research question, design the main concept; draft; RN and DDE assisted with the formulation of the research question, wrote the manuscript in consultation with YAJ; YAJ, RN and DDE performed the research, analysis and interpretation of the data; LL and HKJ commented on the manuscript drafts. All authors read and approved the final manuscript.

Acknowledgments

The authors acknowledge the contributions of Siti Solichatul Makkiyyah from the Medical Faculty, Universitas Islam Indonesia and Amanda Safira Dea Hertika from General Practitioner Department, Dr. Sardjito General Hospital for the reviewing and editing of the protocol.

References

1. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016;82(March):115–21.

2. Cho NH, Shaw JE, Karuranga S, Huang Y, Rocha JD, Ohlrogge AW, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract [Internet]. 2018;138:271–81. Available from: https://doi.org/10.1016/j.diabres.2018.02.023

3. Fowler MJ. Diabetes Foundation: microvascular and macrovascular complications of diabetes. Clin Diab. 2008;26(2):77–82.

4. Khetan AK, Rajagopalan S. Prediabetes. Can J Cardiol. 2018;20(4):183–8.

5. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. Lancet. 2012;379(9833):2279–90.

6. Care D. Introduction: Standards of medical care in diabetes 2019. Diabetes Care. 2019;42(Supplement 1):S1–2.

7. Gerstein HC, Santaguida P, Raina P, Morrison KM, Balion C, Hunt D, et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract. 2007;78(3):305–12.

8. Barber SR, Davies MJ, Khunti K, Gray LJ. Risk assessment tools for detecting those with pre-diabetes: a systematic review. Diabetes Res Clin Pract [Internet]. 2014;105(1):1–13. Available from: http://dx.doi.org/10.1016/j.diabres.2014.03.007

9. Khunti K, Gillies CL, Taub NA, Mostafa SA, Hiles SL, Abrams KR, et al. A comparison of cost per case detected of screening strategies for Type 2 diabetes and impaired glucose regulation: modelling study. Diabetes Res Clin Pract. 2012;97(3):505–13.

10. Ouyang P, Guo X, Shen Y, Lu N, Ma C. A simple score model to assess prediabetes risk status based on the medical examination data. Can J Diabetes. 2016;40(5):419–23.

11. Poltavskiy E, Kim DJ, Bang H. Comparison of screening scores for diabetes and prediabetes. Diabetes Res Clin Pract. 2016;118:146–53.

12. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, et al. Screening for prediabetes using machine learning models. Hindawi Publ. 2014;1–8.

13. Chung JW, Kim WJ, Beom Choi S, Park JS, Kim DW. Screening for prediabetes using support vector machine model. 2014 36th Annu Int Conf IEEE Eng Med Biol Soc. 2014;2472–5.

14. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Informatics. 2019;1–5.

15. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.

16. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P SL. Prisma-P 2015. Br Med J. 2015;349:g7647.

17. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. Manual for cohort and case-control studies. Ottawa Hosp Res Inst. 2013.

18. Bradley AE. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145–59.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.docx
- additionalfile2.docx
- Additionalfile3.docx