





Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis

Satoru Kodama^{1,2} , Kazuya Fujihara^{2*}, Chika Horikawa³, Masaru Kitazawa¹, Midori Iwanaga^{1,2}, Kiminori Kato^{1,2}, Kenichi Watanabe^{1,2}, Yoshimi Nakagawa⁴ , Takashi Matsuzaka⁵ , Hitoshi Shimano⁵, Hirohito Sone² 

¹Department of Prevention of Noncommunicable Diseases and Promotion of Health Checkup, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan, ²Department of Hematology, Endocrinology and Metabolism, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan, ³Department of Health and Nutrition, Faculty of Human Life Studies, University of Niigata Prefecture, Niigata, Japan, ⁴Division of Complex Biosystem Research, Institute of Natural Medicine, Toyama University, Toyama, Japan, and ⁵Department of Internal Medicine (Endocrinology and Metabolism), Faculty of Medicine, University of Tsukuba, Ibaraki, Japan

Keywords

Machine learning, Meta-analysis, Type 2 diabetes mellitus

*Correspondence

Kazuya Fujihara
Tel: +81-25-227-2117
Fax: +81-25-227-2117
E-mail address:
kafujihara-dm@umin.ac.jp

J Diabetes Investig 2022; 13: 900–908

doi: 10.1111/jdi.13736

ABSTRACT

Aims/Introduction: Recently, an increasing number of cohort studies have suggested using machine learning (ML) to predict type 2 diabetes mellitus. However, its predictive ability remains inconclusive. This meta-analysis evaluated the current ability of ML algorithms for predicting incident type 2 diabetes mellitus.

Materials and Methods: We systematically searched longitudinal studies published from 1 January 1950 to 17 May 2020 using MEDLINE and EMBASE. Included studies had to compare ML's classification with the actual incidence of type 2 diabetes mellitus, and present data on the number of true positives, false positives, true negatives and false negatives. The dataset for these four values was pooled with a hierarchical summary receiver operating characteristic and a bivariate random effects model.

Results: There were 12 eligible studies. The pooled sensitivity, specificity, positive likelihood ratio and negative likelihood ratio were 0.81 (95% confidence interval [CI] 0.67–0.90), 0.82 [95% CI 0.74–0.88], 4.55 [95% CI 3.07–6.75] and 0.23 [95% CI 0.13–0.42], respectively. The area under the summarized receiver operating characteristic curve was 0.88 (95% CI 0.85–0.91).

Conclusions: Current ML algorithms have sufficient ability to help clinicians determine whether individuals will develop type 2 diabetes mellitus in the future. However, persons should be cautious before changing their attitude toward future diabetes risk after learning the result of the diabetes prediction test using ML algorithms.

INTRODUCTION

The prevalence of type 2 diabetes mellitus constitutes a worldwide epidemic¹. Intensive lifestyle interventions were shown to reduce the risk of type 2 diabetes mellitus². However, providing such costly programs to entire populations is not feasible. Developing accurate methods for predicting type 2 diabetes mellitus is essential to identify individuals at high risk who should be targeted by type 2 diabetes mellitus prevention programs and to avoid burdening low-risk individuals with unnecessary regimens.

Many researchers have proposed type 2 diabetes mellitus predictive models that typically involve risk scores³. Their

limitations have been low external validation³, time-consuming data collection⁴, a limited set of variables⁵ and potential bias caused by dependence on prior publications for the identification of predictors⁶.

Machine learning (ML) overcomes these weaknesses and has drawn increasing attention in medical research⁷. Evidence to support utilization of ML has been established for identifying diseases, such as melanoma⁸, brain tumors⁹ and sepsis¹⁰. However, evidence is limited for predicting onset of diseases¹¹, possibly because utilization of ML is less prevalent in longitudinal studies compared with cross-sectional studies. Nevertheless, with regard to diabetes, recently an increasing number of cohort studies have tried to use ML for predicting incident diabetes. However, the predictive ability of ML remains

Received 13 July 2021; revised 9 December 2021; accepted 13 December 2021

inconclusive. The aim of the present meta-analysis is to evaluate the current ability of ML algorithms for predicting incident type 2 diabetes mellitus.

MATERIALS AND METHODS

Search strategy

This meta-analysis was reported according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement¹².

Electronic databases (EMBASE and MEDLINE) were used to search for eligible studies published from 1 January 1950 to 17 May 2020. Search terms were thesaurus and text words related to ML, text terms related to type 2 diabetes mellitus, and text terms related to cohort studies. These three elements were combined using the BOOLEAN operator 'AND' (Appendix S1). Inclusion criteria were as follows: (i) longitudinal study; (ii) as the index test, the study used an ML algorithm for determining whether or not each participant will develop type 2 diabetes mellitus in the future; and (iii) as the reference standard, the study ascertained whether or not each participant actually developed type 2 diabetes mellitus during the follow-up period.

One critical characteristic of ML algorithms is a function that operates on input variables to predict a response variable without a hypothesis for a stochastic data model (e.g., Cox regression)¹³. Thus, included studies had to: (i) train two or more possible ML algorithms, and choose one or use an ensemble that combined these algorithms; or (ii) embed the selection of features in the ML algorithm to concurrently build and test a model. Studies meeting neither of those two criteria were excluded.

The outcome of the present meta-analysis was the extent to which results of the index test and reference standard were consistent. Included studies had to present data on the number of true positives, false positives, true negatives and false negatives. If not directly presented, we identified a point maximizing the Youden Index (calculated as [sensitivity + specificity - 1]) from the receiver operating characteristic (ROC) curve, and measured sensitivity and specificity using graphics software (Canvas 11; ACD Systems of America, Inc., Seattle, WA, USA). If one study had two or more ML classification models, we chose the model with the largest area under the ROC.

Data extraction

Two authors (SK and KF) independently extracted data relevant to study characteristics. Disagreements were resolved by a third author (H So). Extracted variables were the first author, published year, location (country), follow-up period, mean age, percentage of men, number of participants and cases, ML classifier, methods for ascertaining type 2 diabetes mellitus, method for separating training and test data, and features that were finally selected for predicting type 2 diabetes mellitus. We used a revised tool for quality assessment of diagnostic accuracy in studies (QUADUS-2) to evaluate study quality. The QUADUS-2 consists of four domains: selection of participants, index test,

reference standard, and flow and timing. All four domains assessed risk of bias, and the first three domains were also used to assess consensus of applicability¹⁴. One question was related to each domain as to the risk of bias and/or applicability for a total of seven questions (Appendix S2). One point was given for each 'yes' answer.

Statistical analysis

We synthesized the dataset consisting of the number of true positives, false positives, true negatives and false negatives in each study. The pooled sensitivity, specificity, positive likelihood ratio (PLR; calculated as [sensitivity / (1 - specificity)]) and negative likelihood ratio (NLR; calculated as [(1 - sensitivity)/specificity]) were estimated by a hierarchical summary ROC¹⁵. The area under the 'summarized' ROC curve (AUROC), wherein sensitivity and specificity in each study were 'summarized', was estimated by a bivariate random effects model. The hierarchical summary ROC and bivariate random effects models have different approaches, but give a consistent result¹⁶. For the PLR and NLR, study heterogeneity was assessed by I^2 ¹⁷ using a multivariate random effects meta-regression, which considered within- and between-study correlations¹⁸. Publication bias was statistically assessed, as proposed by Deeks *et al.*,¹⁹ where a logarithm of the diagnostic odds ratio is regressed against its corresponding inverse of the square root of the effective sample size. Two-sided P value <0.05 was considered statistically significant. All statistical analyses were carried out using STATA 16 (StataCorp., College Station, TX, USA).

RESULTS

Literature searches

Figure 1 shows the procedure for selecting studies for evaluation. There were 1,086 articles retrieved from MEDLINE and/or EMBASE. Among the 13 studies that met our initial inclusion criteria (see Materials and Methods), we had to exclude one study²⁰ from our meta-analysis. That study did not provide reliable information because of inconsistencies between the text and tables, which made it impossible to determine the four values (i.e., number of true positives, false positives, true negatives and false negatives). We queried the author of the study about this serious error, but received no response. Finally, 12 eligible studies^{4-6,21-29} were selected.

Summary of characteristics and quality of the included studies

Study characteristics are described in Table 1. Follow-up duration ranged from 1 to 8 years (median 4.5 years). Included classifiers were decision tree, forward neural network, k-nearest neighbor, logistic regression, logistic regression using the L1 regularization method, support vector machine, random forests, reverse engineering and forward simulation and an ensemble of three decision trees. Selected features varied from six to 1,312.

Table 2 is a summary of selected features from the 12 included studies. Of the 11 representative features, the most frequently selected features were age, obesity and blood glucose

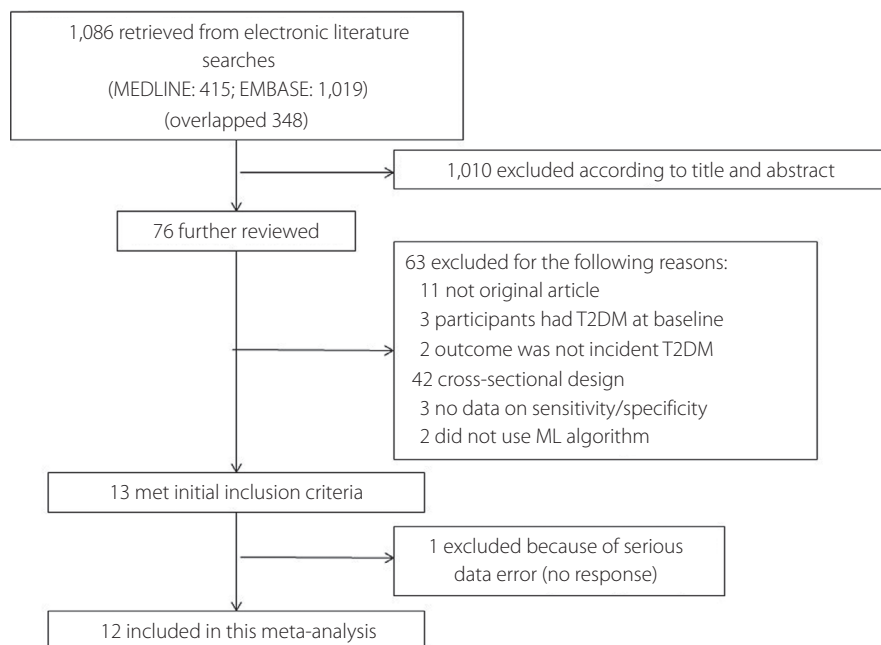


Figure 1 | Study flow in this meta-analysis. ML, machine learning; T2DM, type 2 diabetes mellitus.

(nine studies). Rarely selected features were physical activity (one study) and family history of diabetes (three studies), although these factors are well known as traditional type 2 diabetes mellitus risk factors.

Table 3 shows results of the assessment of study quality using QUADAS-2. The mean score (standard deviation) for study quality was 5.1 (1.0). This low-quality score was mainly attributed to a high risk of bias in the index test in nine studies for at least one of the following reasons: (i) results of the index test were not interpreted without knowledge of the results of the reference standard because of a case-control or historical cohort design; and (2) the threshold of the index test was not pre-specified, because the *n*-fold cross-validation method was used for separating training and test data.

Data synthesis

For each study, the estimated sensitivity, specificity and Youden Index for predicting incident type 2 diabetes mellitus are shown in Table 4. Sensitivity, specificity and the Youden Index in each study ranged from 0.31 to 0.99, from 0.65 to 0.99 and from 0.28 to 0.81, respectively. Based on these values, the hierarchical summary ROC curve and pooled estimates of sensitivity and specificity are shown in Figure 2. The point estimates (95% confidence interval [CI]) of sensitivity, specificity, PLR and NLR were 0.81 (0.67–0.90), 0.82 (0.74–0.88), 4.55 (3.07–6.75) and 0.23 (0.13–0.42), respectively. The AUROC was 0.88 (95% CI 0.85–0.91). Study heterogeneity expressed as I^2 was large (100%, 95% CI 98–100% for PLR and 100%, 95% CI 99–100% for NLR). Publication bias was not statistically significant

($P = 0.99$), which is visually supported by Deek's funnel plot (Appendix S3).

DISCUSSION

While we prepared the present manuscript, a meta-analysis having the same topic as ours was published³⁰. However, that meta-analysis had a significant defect, as its protocol combined longitudinal and cross-sectional studies, which means that the prognostic value was confused with the diagnostic value. In addition, it did show the overall AUROC, but failed to include pooled sensitivity, specificity, PLR and NLR as study end-points that are essential for each individual to predict their incident disease risk³¹.

Equal AUROC does not mean that two ROC curves are identical, although the AUROC is a measure of overall performance. For example, Figure 3 shows two ROC curves, with test A and test B having the same AUROC. However, at the point where the Youden Index (i.e., sensitivity + specificity – 1) is maximized, the PLR and the NLR is $0.96 / (1-0.5) = 1.92$ and $(1-0.96) / 0.5 = 0.08$, respectively, for test A and $0.5 / (1-0.96) = 12.5$ and $(1-0.5) / 0.96 = 0.52$, respectively, for test B. When screening a high-risk group, the test should provide high-sensitivity/low NLR, because otherwise a false negative test could result in many individuals developing a disease that might have been prevented by early interventions³². In contrast, in screening populations with a low prevalence of a disease, high specificity/high PLR is required, because false positives could lead to many unnecessary examinations and treatments³². In the former situation (e.g., Pima Indians in type 2 diabetes

Table 1 | Characteristics of included studies

Study source	Country	Duration (years)	P	C	Age, years (range)	Men (%)	Classifier	Separation†	Methods for ascertaining type 2 diabetes mellitus	No. features
Cahn (2020) ²¹	UK (THIN)	1	137,984	7,878	63	52	DT	HO	MR	69
	Canada (AppleTree)		381,872	17,922	56	49		nCV		
	Israel (MHS)		12,951	693	58	51				
Abbas (2019) ²²	USA	7.5	1,492	171	[25–64]	–	SVM	nCV	Records	4
Choi (2019) ⁴	Korea	5	8,454	404	54	47	LR	nCV	MR	20
Farran (2019) ²³	Kuwait	7	1,837	647	60	50	k-NN	nCV	MR	6
Nguyen (2019) ²⁴	USA	4	9,948	1,890	[21–93]	43	FNN	nCV + HO	RG	1,312
Talaei-Khoei (2018) ⁶	USA	8	8,990	3,073	–	–	SVM	HO	MR	15
Alghamdi (2017) ²⁵	USA	5	32,555	5,099	–	56	ensemble of 3 DTs	nCV	RG/MR	13
Allalou (2016) ²⁶	Canada	2	244	122	34	0	DT	HO	BL	22
Casanova (2016) ⁵	USA	8	3,633	584	53	37	RF	HO	SR/BL	93
Anderson (2015) ²⁷	USA	3.3	24,331	3,765	57	37	REFS	HO	RG	442
Razavian (2015) ²⁸	USA	2	793,153	19,307	48	55	LR with L1	HO	RG/MR/BL	967
Mani (2012) ²⁹	USA	1	2,280	228	–	–	RF	nCV	RG/MR/BL	16

‘–’ indicates that datum was not available. AppleTree, Appletree Medical Group; BL, blood test; C, cases; DT, decision tree; FNN, forward neural network; HO, hold-out; k-NN, k-nearest neighbor; L1 regularization method; LR, logistic regression; MHS, Maccabi Health Services; MR, medical records; nCV, n-fold cross-validation; P, participants; REFS, reverse engineering and forward simulation; RF, random forests; RG, registry; SVM, support vector machine; THIN, The Health Improvement Network. †Method for separating training and test data.

Table 2 | Summary of features selected by machine learning for predicting type 2 diabetes mellitus

Study source	Age	Sex	BP [#1]	Obesity [#2]	PA [#3]	FH of DM	Glucose [#4]	HDL-C/TG	Chol [#5]	Liver [#6]	Kidney [#7]
Cahn (2020) ²¹	†	†		†			†	†	†	†	
Abbas (2019) ²²							†				
Choi (2019) ⁴	†		†				†		†		
Farran (2019) ²³	†	†	†	†		†					
Nguyen (2019) ²⁴	†	†	†	†			†	†	†	†	†
Talaei-Khoei (2018) ⁶	†		†	†		†	†	†			
Alghamdi (2017) ²⁵	†		†	†	†				†		
Allalou (2016) ²⁶		†									
Casanova (2016) ⁵	†	†	†	†		†	†	†	†		†
Anderson (2015) ²⁷	†	†	†	†			†	†	†	†	†
Razavian (2015) ²⁸			†	†			†	†	†	†	†
Mani (2012) ²⁹	†	†	†	†			†	†			†

[#1] Including systolic blood pressure, diastolic blood pressure, hypertension and blood pressure-lowering agents. [#2] Including body mass index and bodyweight, as well as the presence of obesity. [#3] Including sedentary lifestyle. [#4] Including fasting plasma glucose, hemoglobin A1c, and 2-h plasma glucose after a 75-g oral glucose load. [#5] Including total cholesterol, low-density lipoprotein cholesterol, presence of hyperlipidemia and statin use. [#6] Including aspartate aminotransferase and alanine aminotransferase, as well as the presence of liver disease. [#7] Including creatinine and glomerular filtration rate as well as chronic kidney disease. BP, blood pressure; Chol, cholesterol; FH of DM, family history of diabetes mellitus; HDL-C, high density lipoprotein cholesterol; PA, physical activity, TG, triglycerides. †Indicates that the corresponding feature was selected.

mellitus screening³³), test A is superior to test B, whereas test B is superior to test A in the latter situation (e.g., screening of individuals without classic risk factors, such as obesity and parental history of type 2 diabetes mellitus³⁴). In general, according to these calculations, higher sensitivity corresponds to a lower NLR, whereas higher specificity corresponds to a higher PLR³⁵. A higher PLR in a test identifies individuals with a higher likelihood of developing a disease; in other words, it is a better 'rule in' test. Conversely, a lower NLR identifies persons who could reasonably hope not to develop a disease; thus, such a test performs better as a 'rule-out' test³⁵. The prognostic performance of a test should be judged in the context of the specific situation to which it is applied.

For predicting incident type 2 diabetes mellitus, from the meta-analysis of 12 eligible studies, the current ML algorithms were 0.88 of ROC, 4.55 for the PLR and 0.23 for the NLR. According to Hosmer and Lemeshow³⁶, $0.8 \leq \text{AUROC} < 0.9$ is considered to provide excellent discrimination. This would indicate that the current ML algorithms would help clinicians to distinguish individuals at high risk of type 2 diabetes mellitus from low-risk individuals in clinical practice. However, according to the Users' Guide to Medical Literature³¹, the PLR for prognostic tests needs to be ≥ 5 to moderately increase the probability of incident disease, and the NLR should be ≤ 0.2 to moderately decrease the probability of incident disease after taking the index test. It could be interpreted that diabetes prediction tests using ML algorithms did not have sufficient ability to affect the persons' attitude toward future diabetes risk.

We surveyed the predictive ability of previously proposed type 2 diabetes mellitus risk models described by a previous systematic review³. When the survey was limited to cohort studies

that reported every value for sensitivity, specificity and AUROC, the minimum and maximum values were 0.15 and 0.92 for sensitivity, 0.47 and 0.98 for specificity, and 0.11 and 0.65 for the Youden Index. The mean AUROC was 0.79 (Appendix S4). These values suggest that ML algorithms are superior in predicting type 2 diabetes mellitus to these risk models (see Results). In the studies included in the current meta-analysis, data were gathered from uncontrolled observations on complex systems, such as electronic medical records¹³. In addition, information on many well-known risk factors, such as family history of type 2 diabetes mellitus and physical inactivity, were often unavailable (see Table 2). Considering these disadvantages, it is possible that ML algorithms further outperform established risk models than the above comparison suggested. Future research should compare the ability to predict type 2 diabetes mellitus among ML algorithms, previously established risk models and, furthermore, clinicians, using the same database. In addition, using state-of art technology (e.g., Transformer, Gated Recurrent Networks) would make it possible to improve ML performance. Research for predicting future type 2 diabetes mellitus risk should be focused on the further development of ML models considering their high potentiality.

The strength of the present study is that this meta-analysis is the first to assess the current ability of ML to predict incident type 2 diabetes mellitus after a systematic search for longitudinal studies. As shown in Figure 1, there have been many cross-sectional studies of detection of undiagnosed type 2 diabetes mellitus using ML algorithms. The rationale for limiting the analysis to longitudinal studies is that effective methods for identifying individuals who will develop type 2 diabetes mellitus in the future is in greater demand than those for identifying individuals that already have type 2 diabetes mellitus,

Table 3 | Results of assessing study quality using revised tool for the quality assessment of diagnostic accuracy studies (QUADAS-2)

Study source	Risk of bias			D1 score			D2 score			D3 score			D4 score			SQ1	SQ2	SQ3	Applicability score	D1 score	D2 score	D3 score	Total score
	SQ1	SQ2	SQ3	SQ1	SQ2	SQ3	SQ1	SQ2	SQ3	SQ1	SQ2	SQ3	SQ1	SQ2	SQ3								
Cahn (2020) ²¹	Yes	Yes	Yes	1	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	0	1	1	1	6	
Abbas (2019) ²²	Yes	Yes	Yes	1	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	6	
Choi (2019) ⁴	Yes	Yes	Yes	1	No	No	0	No	No	No	0	No	No	No	No	No	No	0	1	1	1	4	
Farran (2019) ²³	Yes	Yes	No	0	No	No	0	No	No	No	0	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	4	
Nguyen (2019) ²⁴	Yes	Yes	Yes	1	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	6	
Talaei-Khoei (2018) ⁶	Yes	Yes	Yes	1	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	0	1	1	1	5	
Alghamdi (2017) ²⁵	No	Yes	Yes	0	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	5	
Casanova (2016) ⁵	Yes	Yes	Yes	1	Yes	Yes	1	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	7	
Allalou (2016) ²⁶	Yes	No	No	0	No	No	0	No	No	No	0	Yes	Yes	Yes	Yes	Yes	Yes	0	1	1	1	4	
Anderson (2015) ²⁷	Yes	Yes	Yes	1	Yes	Yes	1	No	Yes	No	0	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	5	
Razavian (2015) ²⁸	Yes	Yes	Yes	1	Yes	Yes	1	No	Yes	No	0	Yes	Yes	Yes	Yes	Yes	Yes	1	1	1	1	5	
Mani (2012) ²⁹	Yes	No	No	0	No	No	0	No	No	No	0	Yes	Yes	Yes	Yes	Yes	Yes	0	1	1	1	4	

The criterion corresponding to each domain (D) and signaling question (SQ) is indicated in Appendix S2.

Table 4 | Estimated sensitivity and specificity for predicting incident type 2 diabetes using machine learning

Study source	Sensitivity	Specificity	Youden Index
Cahn (2020) ²¹			
THIN	0.82	0.75	0.57
MHS	0.88	0.83	0.70
AppleTree	0.85	0.81	0.66
Abbas (2019) ²²	0.82	0.99	0.80
Choi (2019) ⁴	0.78	0.65	0.43
Farran (2019) ²³	0.72	0.73	0.44
Nguyen (2019) ²⁴	0.31	0.97	0.28
Talaei-Khoei (2018) ⁶	0.68	0.95	0.63
Alghamdi (2017) ²⁵	0.99	0.75	0.74
Allalou (2016) ²⁶	0.74	0.69	0.43
Casanova (2016) ⁵	0.75	0.74	0.49
Anderson (2015) ²⁷	0.66	0.75	0.40
Razavian (2015) ²⁸	0.77	0.70	0.47
Mani (2012) ²⁹	0.76	0.73	0.49

AppleTree, Appletree Medical Group; MHS, Maccabi Health Services; THIN, The Health Improvement Network.

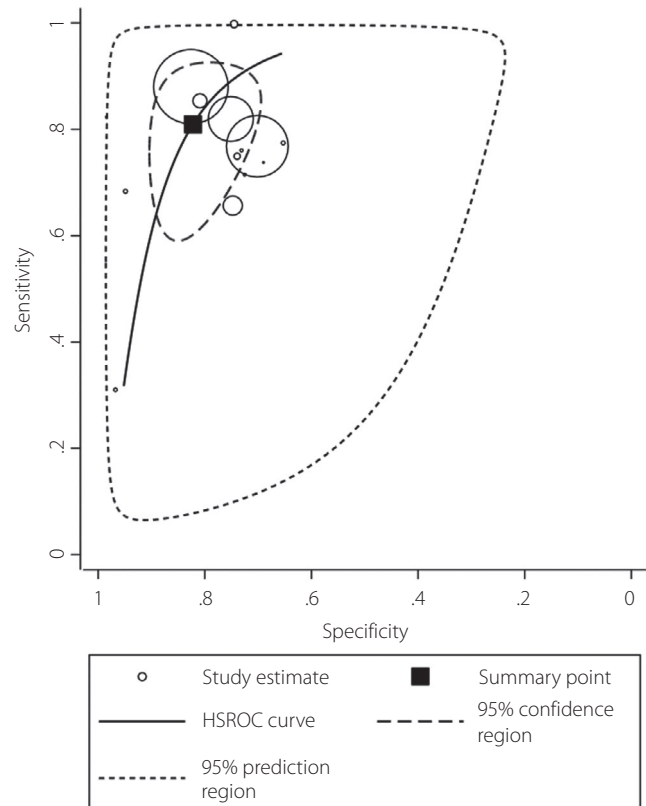


Figure 2 | The hierarchical summary receiver operating characteristic (HSROC) curve for prediction of type 2 diabetes mellitus using machine learning algorithms. The size of each circle is proportional to study sample size. The pooled point estimates of sensitivity and specificity are plotted in a filled square.

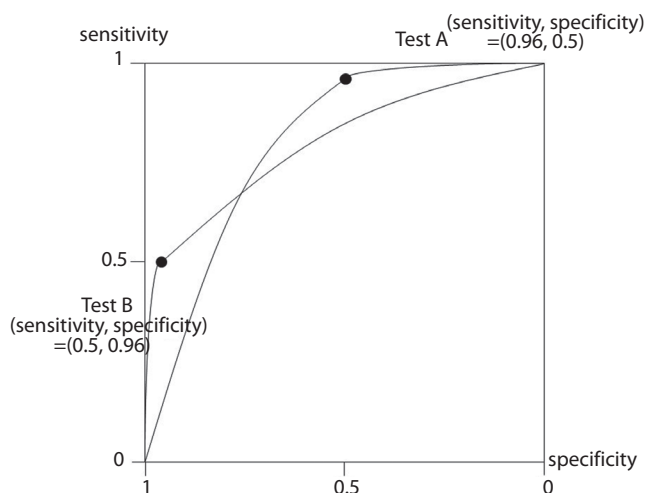


Figure 3 | Illustration of two tests showing receiver operating characteristic curves. Reproduced with permission by Park *et al.*, Korean Journal of Radiology, 2004³² with slight changes. Copyright and all rights reserved.

considering that diabetes treatment causes a substantial economic burden on individuals and health systems¹.

We must address several study limitations. First, one included study in the present meta-analysis²⁴ suggested that deep learning models that have been developed outperform classical ML methods. However, no articles except for that study used such sophisticated models, which might have underestimated the ability of ML to predict type 2 diabetes mellitus. Second, we could use only two search engines (i.e., EMBASE and MEDLINE), because it was technically difficult to use other databases. In addition, the literature searches could not cover state-of-the-art models having specific names, because such names would not be included in the thesaurus terms. The combination of the two search engines is the most optimal in terms of identifying relevant studies³⁷. Furthermore, manual searches were carefully carried out to identify relevant articles that were not retrieved from the literature searches. Although we believe that no eligible study was missed, the existence of an unidentified relevant study cannot be ruled out. Third, studies included in the present meta-analysis were carried out in high-income countries. Evidence of the usefulness of ML in low-to-middle income countries could not be established, although the burden of diabetes treatment is particularly large in such countries.

In conclusion, current ML algorithms are effective resources for clinicians to determine whether an individual will develop type 2 diabetes mellitus in the future. It is likely that the current ML algorithms have already outperformed traditional risk models to predict type 2 diabetes mellitus. However, persons should be cautious regarding changing their attitude toward future diabetes risk after getting results of a diabetes prediction test using ML algorithms. The ML algorithms have a high potential for further improvement of predictive ability for

type 2 diabetes mellitus, although such an improvement might depend on data completeness. Continuous efforts should be made to develop more accurate ML algorithms than currently exist, given that the feasibility of applying ML in a clinical setting would be enhanced in comparison with relying on frequent costly and time-consuming blood tests.

ACKNOWLEDGMENTS

All authors thank Ms Haga and Ms Tada in the Niigata University for their excellent secretarial work. Satoru Kodama was financially supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS; ID:19K12840). The sponsor had no influence over the design and conduct of the study; collection, management, analysis and interpretation of the data; or preparation, review or approval of the manuscript.

DISCLOSURE

The authors declare no conflict of interest.

Approval of the research protocol: N/A.

Informed consent: N/A.

Registry and the registration no. of the study/trial: We registered our protocol in PROSPERO (ID: CRD420201636821; date: 28 April 2020).

Animal studies: N/A.

REFERENCES

1. Roglic G. WHO Global report on diabetes: a summary. *Int J Non-Commun Dis* 2016; 1: 3.
2. Gillies CL, Abrams KR, Lambert PC, *et al.* Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. *BMJ* 2007; 334: 299.
3. Noble D, Mathur R, Dent T, *et al.* Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011; 343: d7163.
4. Choi BG, Rha SW, Kim SW, *et al.* Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med J* 2019; 60: 191–199.
5. Casanova R, Saldana S, Simpson SL, *et al.* Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. *PLoS One* 2016; 11: e0163942.
6. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: a comparison of predictive analytics techniques and predictor variables. *Int J Med Inform* 2018; 119: 22–38.
7. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001; 23: 89–109.
8. Rajpara SM, Botello AP, Townend J, *et al.* Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. *Br J Dermatol* 2009; 161: 591–604.

9. Nguyen AV, Blears EE, Ross E, *et al.* Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. *Neurosurg Focus* 2018; 45: E5.
10. Islam MM, Nasrin T, Walther BA, *et al.* Prediction of sepsis patients using machine learning approach: a meta-analysis. *Comput Methods Programs Biomed* 2019; 170: 1–9.
11. Lee Y, Ragguett R-M, Mansur RB, *et al.* Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord* 2018; 241: 519–532.
12. Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* 2009; 6: e1000097.
13. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* 2001; 16: 199–231.
14. Whiting PF, Rutjes AW, Westwood ME, *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155: 529–536.
15. Harbord R, Whiting P. Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata J* 2009; 9: 211–229.
16. Harbord RM, Deeks JJ, Egger M, *et al.* A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; 8: 239–251.
17. Higgins JP, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003; 327: 557–560.
18. Multivariate WI, Meta-regression R-E. Multivariate random-effects meta-regression: updates to Mvmeta. *Stata J* 2011; 11: 255–270.
19. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; 58: 882–893.
20. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, *et al.* Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput* 2020; 58: 991–1002.
21. Cahn A, Shoshan A, Sagiv T, *et al.* Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. *Diabetes Metab Res Rev* 2020; 36: e3252.
22. Abbas HT, Alic L, Erraguntla M, *et al.* Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLoS One* 2019; 14: e0219636.
23. Farran B, AlWotayan R, Alkandari H, *et al.* Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: a retrospective cohort study of health data from Kuwait. *Front Endocrinol* 2019; 10: 624.
24. Nguyen BP, Pham HN, Tran H, *et al.* Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed* 2019; 182: 105055.
25. Alghamdi M, Al-Mallah M, Keteyian S, *et al.* Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford Exercise Testing (FIT) project. *PLoS One* 2017; 12: e0179805.
26. Allalou A, Nalla A, Prentice KJ, *et al.* A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes. *Diabetes* 2016; 65: 2529–2539.
27. Anderson JP, Parikh JR, Shenfeld DK, *et al.* Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *J Diabetes Sci Technol* 2015; 10: 6–18.
28. 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: 277–287.
29. Mani S, Chen Y, Elasy T, *et al.* Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc* 2012; 2012: 606–615.
30. Silva K, Lee WK, Forbes A, *et al.* Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. *Int J Med Inform* 2020; 143: 104268.
31. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994; 271: 703–707.
32. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol* 2004; 5: 11–18.
33. Knowler WC, Bennett PH, Hamman RF, *et al.* Diabetes incidence and prevalence in pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am J Epidemiol* 1978; 108: 497–505.
34. 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: 1068–1074.
35. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008; 29(Suppl 1): S83–S87.
36. Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd edn. New York: Wiley, 2000.
37. Bramer WM, Rethlefsen ML, Kleijnen J, *et al.* Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev* 2017; 6: 245.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1 | Search strategy in this meta-analysis.

Appendix S2 | Study quality assessment using the quality assessment of diagnostic accuracy studies (QUADAS-2).

Appendix S3 | Deeks' funnel plot asymmetry test for publication bias.

Appendix S4 | Summary of predictive ability for incident type 2 diabetes using traditional risk models limited to cohort studies that reported every value for sensitivity, specificity and area under the receiver operating characteristic curve (AUROC).³