

1 **Predicting the risks for stroke, cardiovascular disease, and** 2 **peripheral vascular disease among people with type 2 diabetes** 3 **with artificial intelligence models: a systematic review and meta-** 4 **analysis**

5 Aqsha Nur,¹ Sydney Tjandra,² Defin Yumnanisha,² Arnold Keane,² Adang Bachtiar¹

6 ¹ Faculty of Public Health, University of Indonesia.

7 ² Faculty of Medicine, University of Indonesia.

8 Correspondence should be addressed to Sydney Tjandra: sydney.tjandra@ui.ac.id

9 **Abstract**

10 **Objectives:** This systematic review and meta-analysis aim to explore the performance of
11 machine learning algorithms in predicting the risk of macrovascular complications among
12 individuals with T2DM, specifically, the predictive capabilities of AI models in forecasting
13 stroke, CVD, and PVD in LMICs.

14 **Design:** Systematic review and meta-analysis of studies reporting on AI prediction models
15 for macrovascular complications in T2DM patients.

16 **Setting:** The review included studies conducted in various healthcare settings, primarily from
17 LMICs, upper-middle-income countries (UMICs), and high-income countries (HICs).

18 **Participants:** 46 studies were included, with a total of 184 AI models. Participants were
19 diverse in age, sex, and geographical locations, reflecting a broad range of healthcare
20 settings.

21 **Interventions:** The intervention analyzed was the application of AI models, including
22 machine learning algorithms, to predict macrovascular complications such as stroke, CVD,
23 and PVD.

24 **Primary and Secondary Outcome Measures:** The primary outcome was the predictive
25 performance of AI models, measured by the area under the receiver operating characteristic
26 curve (AUROC). Secondary outcomes included subgroup analyses based on predictor types
27 and an assessment of AI model applicability in low-resource settings.

28 **Results:** Twelve included studies yielded 184 AI models with an overall AUROC of 0.753
29 (95%CI: 0.74-0.766; I2=99.99%; p<0.001). For 80 models of cardiovascular outcomes, an
30 AUROC of 0.741 (95%CI: 0.721-0.76; I2=99.78%; p<0.001) was obtained. Meanwhile, 25
31 models of peripheral vascular disease and 38 models of cerebrovascular diseases obtained
32 AUROCs of 0.794 (95%CI: 0.758-0.831; I2=97.23%; p<0.001) and 0.77 (95%CI: 0.743-
33 0.797; I2=99.73%; p<0.001) respectively. Subgroup analysis revealed that models with lab-
34 only predictors were superior to those with mixed or no-lab predictors. This signalled the lack

35 of AI capability for history-taking and physical examination data alone, primarily available in
36 low-resource settings.

37 **Conclusions:** Artificial intelligence is promising in predicting diabetes complications.
38 Nevertheless, future studies should explore accessible features in low-resource settings and
39 employ external validation to ensure the robustness of the prediction models.

40 **Keywords:** *artificial intelligence; Diabetes Mellitus, Type 2; Stroke; Cardiovascular*
41 *Disease; Diabetic nephropathy & vascular disease;*

42 **Word Count:** 3685 words

43 **Article Summary**

44 **Strengths and limitations of this study:**

- 45 • Inclusion of studies from both health-related and computer science databases (such as
46 IEEE Xplore) ensured a comprehensive assessment of AI models for predicting
47 diabetes complications.
- 48 • The study analyzed a wide range of models from various countries with different
49 income levels, enhancing the generalizability of the findings.
- 50 • Detailed subgroup analyses provided insights into the impact of predictor types (lab
51 vs. non-lab) and machine learning algorithms on model performance.
- 52 • High heterogeneity across studies, stemming from variations in populations, data
53 sources, and algorithms, was observed, reflecting a common issue in AI model
54 performance meta-analyses.
- 55 • A significant limitation was the lack of external validation in most included studies,
56 which raises concerns about the generalizability and applicability of the AI models in
57 diverse clinical settings.

58 **Key Messages**

- 59 • **What is already known on this topic**
 - 60 ○ Artificial intelligence (AI) holds great potential for diabetes care.
 - 61 ○ Previous meta-analyses have shown its promise in diabetes predictions, but
62 none has been done for diabetes complication predictions.
- 63 • **What this study adds**
 - 64 ○ AI model performance aggregates provided promising results.
 - 65 ○ Subgroup analyses exposed characteristics facilitating prediction
66 performances, namely gradient-boosting algorithms, lab predictors, cross-
67 validation, and detailed missing data.
- 68 • **How this study might affect research, practice, or policy**
 - 69 ○ Albeit promising, ethical open-source models enabling multiple external
70 validations and interdisciplinary collaboration are vital before broader
71 implementation.

72 **Introduction**

73 At least 500 million people were estimated to live with diabetes in 2021, of which 96% were
74 type 2 diabetes mellitus (T2DM).¹ Diabetes complications, such as stroke, cardiovascular
75 diseases (CVD), and peripheral vascular diseases (PVD), increase the 5-year mortality,
76 particularly for people living in low- and middle-income countries (LMICs).² According to
77 World Health Organization (WHO) estimates, 75% of CVD deaths occur in LMICs.³ As the
78 global burden continues to rise, there is an urgent need for precise and early risk stratification
79 methods to enable timely preventive measures for T2DM complications.⁴ In this context, the
80 use of artificial intelligence (AI) and machine learning (ML) models has garnered significant
81 interest for their potential to enhance predictive accuracy in the management of T2DM
82 complications.^{5,6} These technologies promise to transform traditional healthcare approaches
83 by leveraging vast amounts of data to uncover complex patterns and relationships that may
84 not be readily apparent through conventional statistical methods.⁷

85 Previous systematic reviews have primarily focused on the potential of AI in various aspects
86 of diabetes care, particularly in predicting the onset of diabetes itself. For instance, recent
87 meta-analyses have demonstrated the utility of AI in forecasting diabetes-related outcomes,
88 yet none have comprehensively addressed the prediction of macrovascular complications
89 associated explicitly with T2DM.⁸⁻¹⁰ This gap highlights the necessity of a focused
90 investigation into how AI can be harnessed to predict complications like stroke, CVD, and
91 PVD in patients already diagnosed with T2DM, including its deployment in low-resource
92 settings.¹¹

93 This systematic review and meta-analysis aim to expdialabore the performance of machine
94 learning algorithms in predicting the risk of macrovascular complications among individuals
95 with T2DM, specifically, the predictive capabilities of AI models in forecasting stroke, CVD,
96 and PVD. Unlike earlier studies that may have concentrated on a single type of complication
97 or generalized diabetes prediction, this review delves into a broader spectrum. Moreover, it
98 provides an in-depth analysis of subgroup performances, comparing models with various
99 predictor types, including lab-only and mixed predictors, and examining the implications of
100 these differences. This review also highlights the challenges and limitations associated with
101 current AI models, particularly their applicability in low-resource settings. By focusing on
102 models that utilize widely available data and require minimal specialized input, the findings
103 might guide future research or policy-making for AI tools that can be deployed effectively in
104 regions with limited healthcare infrastructure.

105 **Materials and Methods**

106 **Search Strategy**

107 This review was systematically developed, conducted, and reported following Preferred
108 Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist during
109 writing our report as presented in the Supplementary Material 1.¹² Our protocol has been
110 registered at The International Prospective Register of Systematic Reviews (PROSPERO)
111 under the reference ID CRD42023489167.

112 We searched six databases (Scopus, PubMed, Embase, Wiley Online Library, IEEE Explore,
113 and Google Scholar) and hand-searched for articles published between January 1, 2000, and
114 November 30, 2023. Keywords employed were “type 2 diabetes”, “artificial intelligence,”

115 “prediction,” “complication,” “stroke,” “cardiovascular disease,” and “peripheral vascular
116 disease,” as well as their MeSH terms and subsets combined with Boolean operators (see
117 Supplementary Material 2). Search results were exported and deduplicated to Rayyan
118 (www.rayyan.ai).

119 **Eligibility Criteria**

120 Each article was screened for the following PICOT inclusion criteria (see Supplementary
121 Material 3) by at least two members independently (AN, ST, RH, SW):¹³ (1) subjects are
122 adults aged 18 years old or above with type 2 diabetes mellitus, (2) intervention was the
123 development and implementation of artificial intelligence, including but not limited to
124 machine learning and deep learning, as opposed to classical statistical models, (3) outcome
125 included prediction performances for stroke, cardiovascular disease, or peripheral vascular
126 disease, (4) diagnostic or prognostic studies with a cohort or case-control design capable of
127 exhibiting temporality, (5) used any actual medical dataset, and (6) published in English. We
128 excluded studies that (1) had mixed populations with type 1 and/or prediabetes patients, (2)
129 mainly explained theoretical models not tested on human subjects, (3) involved drugs as the
130 intervention, (4) were reviews, framework developments, conference abstracts, proposals,
131 editorials, commentaries, and qualitative studies, and (5) had irretrievable full-text. After
132 titles and abstracts were screened on Rayyan, full-text screening was conducted to reconfirm
133 eligibility. Discrepancies were resolved through consensus.

134 **Data Extraction**

135 A data extraction instrument was developed to tribulate several characteristics and details
136 from all included studies, namely (1) author and year, (2) country of origin, (3) study design,
137 (4) data source, (5) single or multi-centred, (6) population profile (including number of
138 patients, age, and proportion of males), (7) predictors, (8) whether external validation was
139 employed, (9) AI/ML algorithm, (10) outcome (stroke, cardiovascular disease, or peripheral
140 vascular disease), (11) data period and follow-up, (12) data pre-processing details, and (13)
141 internal validation setup. We also extracted the main outcome, model performance, in metrics
142 such as F-measures, the area under the receiving operating curve (AUROC), c-statistics,
143 sensitivity/recall, specificity, accuracy, and precision/positive predictive value.

144 **Risk of Bias Assessment**

145 Two members (AN, ST, DY, AK) independently assessed all included studies for risk of bias
146 and applicability using the signalling questions on the Prediction Model Risk of Bias
147 Assessment Tool (PROBAST).¹⁴ Differences were discussed to reach a consensus.

148 **Quantitative Data Analysis**

149 Studies reporting AUROCs as model performances were aggregated through a random-
150 effects meta-analysis with MedCalc, visualized with R. When neither the standard error,
151 range, nor the standard deviation was disclosed, we ran the Hanley and McNeil’s approach¹⁵
152 with R to approximate the standard error based on the AUROC, sample size, and number of
153 complication cases.¹⁵⁻¹⁷ To assess publication bias, funnel plots and Egger’s regression were
154 generated with MedCalc. Moreover, outliers, defined as models whose 95% confidence
155 intervals did not overlap with the meta-analysis result, were excluded to generate sensitivity
156 analyses. As substantial heterogeneity remains, subgroup analyses were conducted for

157 outcome types, external validation, algorithms, country income levels, risk of bias, missing
158 data process details, cross-validation, and predictor data.

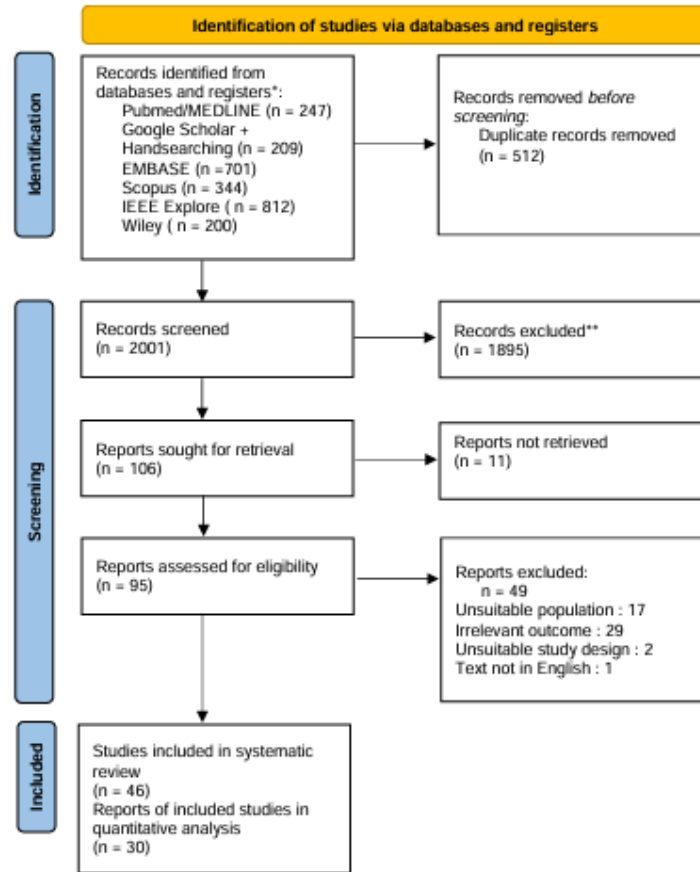
159 **Results and Discussion**

160 **Study Characteristics**

161 A total of 2,513 studies were found during the initial search across seven databases (PubMed,
162 Google Scholar, EMBASE, Scopus, IEEE Explore, and Wiley), in addition to hand
163 searching. After removing 512 duplicate records, 2,001 records were screened for their titles
164 and abstracts. Subsequently, 1,895 records were excluded, leaving 106 reports to be retrieved.
165 Studies without available full texts were excluded, resulting in 95 studies being assessed for
166 eligibility. Of these, 49 studies were excluded for various reasons: unsuitable population (17
167 studies), irrelevant outcome (29 studies), unsuitable study design (2 studies), and text not in
168 English (1 study). Ultimately, 46 studies were included in the systematic review, with 30
169 included in the quantitative analysis. The selection process is depicted in Figure 1.

170 Supplementary Material 4 depicts the overall characteristics of the study, including the
171 participants and outcomes of each study. The systematic review encompasses 46 studies from
172 various countries, which are categorized into three regions based on income levels: low-
173 middle-income countries (LMIC), upper-middle-income countries (UMIC), and high-income
174 countries (HIC). Most data were sourced from hospital medical records in the respective
175 countries, while some datasets were from specific trials or studies. Sample sizes from each
176 study varied from around a hundred to hundreds of thousands, even surpassing a million. The
177 predictors were divided into demographic, clinical, comorbidity, and laboratory groups. The
178 algorithms were classified into several main categories, such as Random Forest, Neural
179 Network, Gradient Boosting, Logistic Regression, Multi-task Learning, Cox-based methods,
180 and others. The outcomes assessed included cardiovascular disease, cerebrovascular disease
181 (stroke), and peripheral vascular disease. Information about technical aspects such as
182 handling missing data, cross-validation, and external validation is also provided.

183



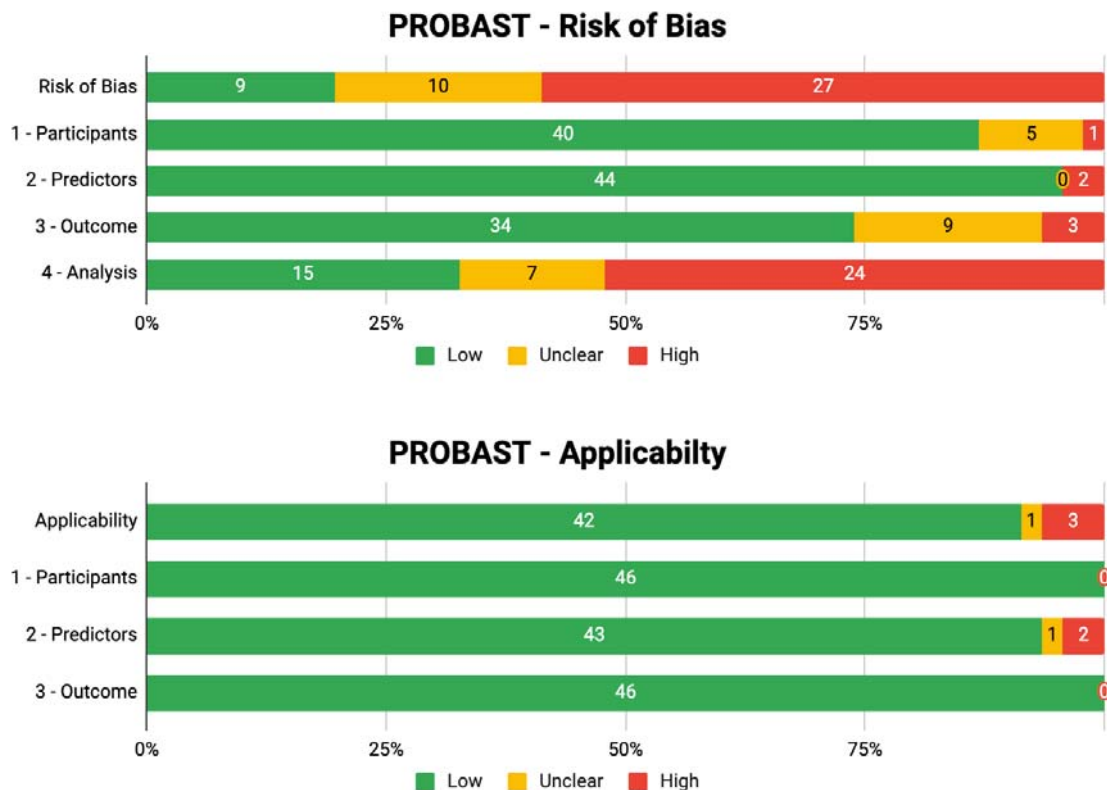
184

185

Figure 1: PRISMA flowchart of included studies.

186 Risk of Bias

187 The overall risk of bias, as summarised in Figure 2, indicates that 78% of the articles are
188 rated as having a high or uncertain risk of bias, with specific distributions of 10 articles rated
189 as low risk, ten as unclear risk, and 27 as high risk. The high risk of bias predominantly
190 originated from the “outcome” domain due to the uncertainty in determining outcomes
191 without knowledge of predictor information. Additionally, the “analysis” domain contributed
192 significantly to the high risk, primarily due to inadequate handling of missing data or
193 improper imputation methods and the low number of participants with the outcome. Nearly
194 all studies (91%) showed no concerns regarding applicability concerns.



195

196

197

Figure 2: PROBABT summary of included studies

198 **Meta-Analysis of AUROCs**

199 All 184 models were pooled with a random effects model, obtaining an AUROC of 0.753
 200 (95%CI: 0.74–0.766; $I^2=99.99\%$; $p<0.001$), as shown through the forest plot in Figure 3. For
 201 80 models of cardiovascular outcomes, an AUROC of 0.741 (95%CI: 0.721–0.76;
 202 $I^2=99.78\%$; $p<0.001$) was obtained. Meanwhile, 25 models of peripheral vascular disease and
 203 38 models of cerebrovascular diseases obtained AUROCs of 0.794 (95%CI: 0.758–0.831;
 204 $I^2=97.23\%$; $p<0.001$) and 0.77 (95%CI: 0.743–0.797; $I^2=99.73\%$; $p<0.001$) respectively
 205 (Supplementary Materials 6–9). Subgroup analysis results are detailed in Table 1.



206

207 Figure 3: Forest plot of artificial intelligence areas under the operating curve (AUROCs) in predicting diabetes
 208 complications.

209 Table 1: Subgroup analyses summary of Area Under the Receiver Operating Characteristics (AUROCs) of
 210 machine learning prediction models for diabetes macrovascular complications based on various characteristics.

Subgroups	No. of prediction models	Random-effects AUROC	Lower 95% CI	Higher 95% CI	Heterogeneity (%)	p-value
All studies	184	0.753	0.74	0.766	99.99	< 0.001
Outcome types						
Cardiovascular	80	0.741	0.721	0.76	99.78	< 0.001
Peripheral vascular/diabetic foot	25	0.794	0.758	0.831	97.23	< 0.001
Stroke / cerebrovascular	38	0.77	0.743	0.797	99.73	< 0.001
Mixed	41	0.741	0.716	0.765	100	< 0.001
External validation data						
Yes	56	0.725	0.708	0.742	97.85	< 0.001
No	128	0.765	0.749	0.782	99.99	< 0.001
Machine learning algorithm						
Cox-based	14	0.712	0.664	0.76	98.53	< 0.001
Gradient boosting	37	0.789	0.761	0.817	99.56	< 0.001
Logistic regression	23	0.731	0.711	0.752	99.55	< 0.001
Multi-task learning	18	0.699	0.665	0.733	99.99	< 0.001
Neural network	11	0.759	0.722	0.797	98.55	< 0.001
Random forest	30	0.776	0.742	0.810	99.73	< 0.001
Others	51	0.752	0.726	0.777	99.99	< 0.001
Country income						
HIC	134	0.737	0.723	0.751	99.99	< 0.001
LIC/LMIC/UMIC	50	0.8	0.774	0.825	97.31	< 0.001
Risk of bias						
Low/medium	110	0.780	0.765	0.794	99.76	< 0.001
High	74	0.711	0.691	0.731	99.99	< 0.001
Missing data process detailed						
Yes	114	0.775	0.76	0.79	99.66	< 0.001
No	70	0.717	0.696	0.738	99.99	< 0.001
Cross-validation						

Yes	127	0.759	0.743	0.775	99.99	< 0.001
No	57	0.739	0.717	0.761	98.88	< 0.001
Predictor data						
No lab	29	0.714	0.696	0.731	100	< 0.001
Lab only	3	0.837	0.784	0.89	0	< 0.001
Mixed	152	0.759	0.745	0.774	99.98	< 0.001

AUROC, Area Under the Receiver Operating Characteristic; CI, confidence interval; HIC, high-income countries; LIC, low-income countries; LMIC, lower-middle-income countries; UMIC, upper-middle-income countries

211 Publication Bias

212 We observed significant publication bias (Egger's test p-value = 0.0261). Funnel plots of
213 AUROCs against standard errors are presented in Supplementary Materials 10—14.

214 Sensitivity Analyses

215 We excluded outliers and retrieved 83 models with an overall AUROC of 0.746 (95%CI:
216 0.742–0.75; $I^2=99.86\%$; $p<0.001$). This is comparable to the initial meta-analysis, showing
217 robustness despite outliers. Similarly, outcome and predictor subgroup sensitivity analyses
218 were conducted, with results in Table 2. Most notably, the peripheral vascular disease
219 outcome subgroup retrieved an AUROC of 0.820 (95%CI: 0.798–0.842; $p<0.001$) with a
220 heterogeneity of $I^2=0\%$. In the lab-only predictors subgroup, no outliers were identified.

221 Table 2: Sensitivity analyses summary of Area Under the Receiver Operating Characteristics (AUROCs) of
222 machine learning prediction models for diabetes macrovascular complications based on outcomes and
223 predictors.

Subgroups	No. of prediction models	Random-effects AUROC	Lower 95% CI	Higher 95% CI	Heterogeneity (%)	p-value
All studies	83	0.746	0.742	0.75	99.86	< 0.001
Outcome types						
Cardiovascular	38	0.741	0.733	0.749	80.99	< 0.001
Peripheral vascular/diabetic foot	15	0.820	0.798	0.842	0.00	< 0.001
Stroke / cerebrovascular	18	0.756	0.747	0.764	92.42	< 0.001
Mixed	16	0.737	0.729	0.745	99.97	< 0.001
External validation data						
No lab	15	0.710	0.703	0.718	99.90	< 0.001

Mixed	73	0.753	0.748	0.758	92.31	< 0.001
-------	----	-------	-------	-------	-------	---------

224 **Discussion**

225 **Model Performance**

226 The pooled analysis of 184 models revealed a moderate level of performance, with an overall
227 AUROC of 0.753. The models demonstrated varying performance based on the specific
228 outcome types. The models achieved an AUROC of 0.741 for cardiovascular disease
229 outcomes, while those predicting peripheral vascular disease and cerebrovascular disease had
230 higher AUROCs of 0.794 and 0.77, respectively. Nanda R et al. (2022) generated an RF
231 model with the highest AUC across all models of 0.918 to predict the risk of T2DM people
232 developing diabetic foot ulcers.¹⁸ These results suggest that while the models are generally
233 robust, their effectiveness can vary depending on the specific type of predicted macrovascular
234 complication.

235 Heterogeneity among the included studies was notably high, with I^2 values approaching
236 100% across most analyses. This substantial heterogeneity underscores the variability in
237 model performance, which could stem from differences in study populations, data sources,
238 predictor variables, and machine learning algorithms used. The high heterogeneity highlights
239 the importance of context-specific factors in model performance and suggests that predictive
240 accuracy may improve when models are tailored to specific populations and settings. The
241 sensitivity analysis further supports the robustness of the findings. By excluding outliers, the
242 overall AUROC was slightly reduced to 0.746, with heterogeneity remaining high
243 ($I^2=99.86\%$). This consistency indicates that extreme values do not unduly influence the
244 overall conclusions.

245 Using these ML models for diabetes complication risk prediction might be helpful in
246 considering several limitations in several existing conventional scoring systems. For
247 example, the Framingham risk score, the most established risk assessment for heart disease,
248 was developed for the general population and not specific for T2DM people.¹⁹ This risk score
249 is designed for the general population and not specifically for people with diabetes. Risk
250 scores developed for general populations may have lower discriminatory ability in
251 individuals with diabetes.²⁰ Other researchers have also developed heart disease risk
252 assessment focus using ML models,²¹ with linear and logistic regression, and artificial neural
253 networks (ANN) often used due to their simplicity and good predictive ability.

254 **Study settings**

255 Most studies (29; 63%) in our review came from research in HICs, followed by UMICs (9;
256 19.6%) and LMICs (8; 17.4%). India accounts for the majority of studies from the LMICs,
257 China dominates the UMICs, and the United States leads in the HICs. This might
258 demonstrate that a country's income level influences the number of artificial intelligence
259 research and publications. A bibliometric study by Jimma (2023) mapped the publication of
260 artificial intelligence in health care.²² Interestingly, the United States and China are included in
261 the top nine countries, with the United States ranking first (41.84%) and China second
262 (14.70%). Our study also identified that the most productive and prominent institutions
263 funding AI research are from the United States, including the National Institutes of Health
264 and the US National Library of Medicine. The disparity in the number of studies in non-high-

265 income countries is due to limited healthcare resources. This is significant considering that
266 80% of the global population resides in developing countries, where public health issues
267 continue to rise due to rapid globalization and urbanization. Therefore, studies in developing
268 countries are crucial, as the lack of data in these regions affects the applicability of findings
269 to their specific contexts. Our study also identified that the most productive and prominent
270 institutions funding AI research are from the United States, including the National Institutes
271 of Health and the US National Library of Medicine. The disparity in the number of studies in
272 non-high-income countries is due to limited healthcare resources. This is significant
273 considering that 80% of the global population resides in developing countries, where public
274 health issues continue to rise due to rapid globalization and urbanization. Therefore, studies
275 in developing countries are crucial, as the lack of data in these regions affects the
276 applicability of findings to their specific contexts.

277 **Predictors**

278 Our study collected 250 different predictors from the 46 studies and grouped them into four
279 categories: demographic (13 predictors; 5.2%), clinical (50 predictors; 20%), comorbidity (33
280 predictors; 13.2%), and laboratory (154 predictors; 61.6%). 42 (91.3%) studies included
281 demographics, with the most used predictors being age, sex, and race. Demographics were
282 included in 42 (91.3%) studies, with age, sex, and race being the most commonly used
283 predictors. Clinical factors were included in 43 (93.5%) studies, featuring predictors like
284 body mass index, blood pressure, and history of antidiabetic medication. Comorbidities were
285 considered in only 23 (50%) studies, including hypertension, heart disease, and renal
286 diseases. Laboratory parameters were utilized in 34 (73.9%) studies, with the most frequent
287 predictors being HbA1c, high-density lipids, and cholesterol levels.

288 Testing an existing ML model in other settings needs to account for the availability of
289 predictor data. In low-resource settings, a model which requires laboratory parameters might
290 not be difficult to test due to limited infrastructure. We created a subgroup analysis of models
291 with no laboratory data (i.e., demographic, clinical, or comorbidity), with only laboratory
292 data, and mixed. The AUROCs of lab-only and non-lab models are 0.837 and 0.714
293 respectively. This means non-lab models were comparable and did not perform poorly
294 compared to lab parameters.^{22 23} To improve performance, several strategies can be
295 employed, such as hyperparameter tuning and exploring different algorithms that can
296 optimize the model.²⁴

297 **Model Development**

298 Most (n=29, 63.04%) included studies that utilized k-fold cross-validation as internal
299 validation, similar to previous studies in diabetes risk prediction.²⁵ With this method, the data
300 is divided into k folds of equal size, and the model is subsequently trained and evaluated k
301 times, with each evaluation utilizing a different fold as the test set.²⁵ This method is
302 preferable compared to the hold-out approach as the whole dataset is utilized for
303 development.^{6 25} However, like other internal validation methods (including bootstrapping),
304 optimism should be adjusted for in the final model.⁶

305 Only 21 (45.65%) studies reported how they handled missing data, although disregarding it
306 may lead to imbalances, consequently introducing bias and misleading results. We found that
307 models where missing data handling is described perform better. Reporting is essential as
308 imputing different central tendencies (mean, median, or mode), as missing data could lead to

309 different outcomes for different data distributions.²⁶ More recently, autoencoders and other
310 imputation techniques can more accurately fill in incomplete data.^{27 28} These technologies
311 would be beneficial for data pre-processing prior to AI model developments.

312 **Algorithm Types**

313 Interestingly, our meta-analysis found gradient boosting to be the most common ML
314 algorithm utilized, with a leading AUROC model performance of 0.789, followed by random
315 forests (AUROC 0.776). Boosting algorithms are similar to random forests as they are
316 ensemble learning algorithms, with the advantage of avoiding overfitting.^{29 30} They also work
317 well with categorical and numerical predictors. The third leading algorithm for performance,
318 neural networks (AUROC 0.759), are comparatively less utilized by studies. As they fall in
319 the deep learning category, despite their exceptional performance and capability to capture
320 complex relationships, they are demanding computationally as they require large datasets.³⁰

321 **External Validation**

322 The uniform decrease of model performance when validated in external datasets (AUROC of
323 0.725) compared to internal validations (AUROC of 0.765) proved that development stages
324 tend to overestimate, consistent with previous studies, such as non-AI prognostic model
325 studies,³¹ or AI models for other purposes.^{32 33} Moreover, only 11 (23.91%) of our included
326 studies conducted external validation, despite it being a crucial step in prediction models and
327 prognostic research, providing the capability for clinical impact over different settings.³⁴
328 Contrarily, for some studies, such as those with small non-representative datasets or missing
329 predictors, an external validation may not be worth it.³⁵ Judging the overly optimistic nature
330 of development models, diabetes complication prediction AI models may consider multiple
331 external validations unless they are specifically made for local clinical settings.³⁶

332 **Risk of Bias**

333 An analysis of the risk of bias in published studies reveals significant issues in their design.
334 First, many studies exhibit a high or unclear risk of bias, often due to incomplete data and
335 insufficient population samplings, such as the underrepresentation of diverse patient groups
336 and inadequate consideration of critical predictors like age and laboratory results—issues
337 with data extraction, including incomplete or inconsistent datasets, further compromise
338 model accuracy and reliability. Variable follow-up intervals also affect the generalizability of
339 results.³⁷

340 Second, The heavy reliance on internal validation with limited datasets from single centres is
341 another concern, as studies lacking external validation show a higher risk of bias. In contrast,
342 multi-center studies or those using national databases tend to have lower risk.³⁸ The future
343 success of machine learning prediction models hinges on high-quality, diverse training data.
344 Proper data handling, capturing heterogeneity, and incorporating complexity are essential to
345 enhance the models' applicability and reliability.³⁹

346 **Way Forward**

347 The application of AI and machine learning (ML) in predicting complications is in its early
348 stages, with significant potential due to the increasing complexity and volume of data. Early
349 and accurate diagnosis of macrovascular complications could enable timely treatment, but

350 this requires rigorous validation and scrutiny for effective outcomes. Enhancing reliability
351 involves increasing external validation from diverse sources and promoting open-source
352 development and interdisciplinary collaboration. While laboratory data enhances predictive
353 accuracy, reliance on such data may limit the applicability of AI models in low-resource
354 settings where lab facilities are not readily available. Future research should aim to improve
355 the predictive power of non-lab models by incorporating advanced techniques for history-
356 taking and physical examination data.

357 To ensure ethical and practical AI/ML use in healthcare, it is crucial to establish a secure
358 framework focusing on data protection, secure handling, patient consent, and algorithmic
359 transparency. Addressing biases and limitations is essential for broader implementation.
360 Collaboration with policymakers, bioethicists, academics, and the broader community will be
361 vital. Future research should prioritize validation and implementation strategies to improve
362 the practical utility and trustworthiness of AI/ML models in clinical settings. Moreover,
363 training healthcare professionals to interpret AI-driven predictions and incorporate them into
364 patient management plans will further enhance the practical utility of these tools. Finally,
365 continuous monitoring and updating of AI models with new data will ensure their ongoing
366 accuracy and relevance in a rapidly evolving healthcare landscape.

367

368 **Strengths and Limitations**

369 With the numerous models included in the meta-analysis, we are confident that this study
370 reflects the capability of artificial intelligence in predicting diabetes complications to date.
371 Information technology literature, such as IEEE Xplore, yielded studies from computer
372 science fields that would have been absent in health-related databases. The included studies
373 were done in multiple countries of varying income levels. Additionally, the detailed subgroup
374 analysis provides valuable insights into the factors affecting model performance, such as the
375 type of predictors (lab vs. non-lab) and the machine learning algorithms used.

376 Nevertheless, despite all relevant subgroup analyses explored, our meta-analysis has high
377 heterogeneity, which can stem from differences in study populations, data sources, and
378 machine learning algorithms used; such a phenomenon is commonly observed in published
379 AI model performance meta-analyses.^{33 40} Only a tiny proportion of the included studies
380 conducted external validation, a crucial step for assessing the generalizability of prediction
381 models. This lack of external validation raises concerns about the models' applicability in
382 different clinical settings. We also only analyzed AUROCs in our meta-analyses as the most
383 utilized parameter; consequently, we excluded studies using different model performance
384 parameters. Finally, the reliance on laboratory data for superior predictive accuracy may limit
385 the practical implementation of these models in low-resource settings, where such data may
386 not be readily available. Future research should focus on enhancing the performance of non-
387 lab-based models to increase their applicability across diverse healthcare environments.

388 **Conclusions**

389 This review demonstrates the promising potential of machine learning (ML) models in
390 predicting macrovascular complications among individuals with type 2 diabetes mellitus
391 (T2DM). We reveal a moderate overall performance with significant insights into the factors
392 influencing predictive accuracy. However, the high heterogeneity observed among the

393 included studies highlights the variability in model performance, emphasizing the need for
394 tailored approaches based on specific populations and settings. Future studies should focus on
395 developing robust non-lab-based models and conducting extensive external validations to
396 improve the applicability of AI models in diverse clinical settings, especially in low-resource
397 environments. Ultimately, the successful integration of AI and ML models in predicting
398 diabetes complications will require interdisciplinary collaboration, ethical considerations, and
399 ongoing validation to ensure their reliability and effectiveness in real-world clinical practice.

400

401 **Author Contributions**

402 **Aqsha Nur:** Conceptualization, review protocol, article screening, data extraction, bias
403 assessment, writing – original draft; **Sydney Tjandra:** Search strategy, article screening, data
404 extraction, bias assessment, quantitative analysis and synthesis, writing – original draft;
405 **Defin Yumnanisha:** data extraction, bias assessment, writing – original draft; **Arnold**
406 **Keane:** data extraction, bias assessment, writing – original draft, writing – review and
407 editing; **Adang Bachtiar:** Conceptualization, review protocol, writing – review.

408 **Conflicts of Interest**

409 The author(s) declare(s) that there is no conflict of interest regarding the publication of this
410 paper.

411 **Funding Statement**

412 None.

413 **Author Contributions**

414 Stevano Wijoyo and Rizqi Humaira (University of Indonesia) supported the initial screening
415 of this review. Dante Harbuwono (University of Indonesia) provided feedback on the clinical
416 impact of AI for diabetes, which guided the protocol of this review.

417 **Acknowledgments**

418 Stevano Wijoyo (University of Indonesia) supported the initial screening of this review.
419 Dante Harbuwono (University of Indonesia) and Sri Laksmiastuti (University of Trisakti)
420 suggested the clinical impact of AI on diabetes, which guided the protocol of this review.

421 **Data Availability Statement**

422 Data of this study are publicly available on the Open Science Framework, a public, open
423 access repository, at <https://osf.io/7gh9m/>. Please contact the corresponding author for any
424 further inquiries.

425 **References**

- 426 1. Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes
427 prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2018;138:271-
428 81. doi: 10.1016/j.diabres.2018.02.023 [published Online First: 20180226]
- 429 2. Raghavan S, Vassy JL, Ho YL, et al. Diabetes Mellitus–Related All-cause and
430 Cardiovascular Mortality in a National Cohort of Adults. *Journal of the American*
431 *Heart Association* 2019;8(4):e011295. doi: doi:10.1161/JAHA.118.011295
- 432 3. World Health Organization. Cardiovascular Diseases (CVDs) Fact Sheet: World Health
433 Organization; [Available from: [https://www.who.int/news-room/fact-](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
434 [sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))] accessed 2 August 2024.
- 435 4. Tomic D, Shaw JE, Magliano DJ. The burden and risks of emerging complications of
436 diabetes mellitus. *Nature Reviews Endocrinology* 2022;18(9):525-39. doi:
437 10.1038/s41574-022-00690-7
- 438 5. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial
439 intelligence. *Nature* 2023;616(7956):259-65. doi: 10.1038/s41586-023-05881-4
- 440 6. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies:
441 Explanation and Elaboration. *Annals of Internal Medicine* 2019;170(1):W1-W33. doi:
442 10.7326/m18-1377 %m 30596876
- 443 7. Hunter DJ, Holmes C. Where Medical Statistics Meets Artificial Intelligence. *New*
444 *England Journal of Medicine* 2023;389(13):1211-19. doi:
445 doi:10.1056/NEJMra2212850
- 446 8. Kee OT, Harun H, Mustafa N, et al. Cardiovascular complications in a diabetes prediction
447 model using machine learning: a systematic review. *Cardiovasc Diabetol*
448 2023;22(1):13. doi: 10.1186/s12933-023-01741-7 [published Online First: 20230119]
- 449 9. Chowdhury MZI, Yeasmin F, Rabi DM, et al. Predicting the risk of stroke among patients
450 with type 2 diabetes: a systematic review and meta-analysis of C-statistics. *BMJ Open*
451 2019;9(8):e025579. doi: 10.1136/bmjopen-2018-025579
- 452 10. Usman TM, Saheed YK, Nsang A, et al. A systematic literature review of machine
453 learning based risk prediction models for diabetic retinopathy progression. *Artificial*
454 *Intelligence in Medicine* 2023;143:102617. doi:
455 <https://doi.org/10.1016/j.artmed.2023.102617>
- 456 11. Ciecierski-Holmes T, Singh R, Axt M, et al. Artificial intelligence for strengthening
457 healthcare systems in low- and middle-income countries: a systematic scoping review.
458 *npj Digital Medicine* 2022;5(1):162. doi: 10.1038/s41746-022-00700-y
- 459 12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated
460 guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi: 10.1136/bmj.n71
461 [published Online First: 20210329]

- 462 13. Riva JJ, Malik KM, Burnie SJ, et al. What is your research question? An introduction to
463 the PICOT format for clinicians. *J Can Chiropr Assoc* 2012;56(3):167-71.
- 464 14. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and
465 Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of*
466 *Internal Medicine* 2019;170(1):W1-W33. doi: 10.7326/M18-1377
- 467 15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating
468 characteristic (ROC) curve. *Radiology* 1982;143(1):29-36. doi:
469 10.1148/radiology.143.1.7063747
- 470 16. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
471 Vienna2016 [cited 2024 1 May 2024]. Available from: <https://www.R-project.org/>.
- 472 17. Gardner J. text: Statistical Testing for AUC Data [Internet]. [cited 2024 1 May 2024].
473 Available from: <https://cran.r-project.org/web/packages/auctestr/auctestr.pdf>.
- 474 18. Nanda R, Nath A, Patel S, Mohapatra E. Machine learning algorithm to evaluate risk
475 factors of diabetic foot ulcers and its severity. *Medical & Biological Engineering &*
476 *Computing* 2022;60(8):2349-57. doi: 10.1007/s11517-022-02617-w
- 477 19. Lloyd-Jones DM, Wilson PW, Larson MG, et al. Framingham risk score and prediction of
478 lifetime risk for coronary heart disease. *Am J Cardiol* 2004;94(1):20-4. doi:
479 10.1016/j.amjcard.2004.03.023
- 480 20. Chamnan P, Simmons RK, Sharp SJ, et al. Cardiovascular risk assessment scores for
481 people with diabetes: a systematic review. *Diabetologia* 2009;52(10):2001-14. doi:
482 10.1007/s00125-009-1454-0 [published Online First: 20090724]
- 483 21. Wang J, Wang S, Zhu MX, et al. Risk Prediction of Major Adverse Cardiovascular
484 Events Occurrence Within 6 Months After Coronary Revascularization: Machine
485 Learning Study. *JMIR Med Inform* 2022;10(4):e33395. doi: 10.2196/33395
486 [published Online First: 20220420]
- 487 22. Nahm FS. Receiver operating characteristic curve: overview and practical use for
488 clinicians. *Korean J Anesthesiol* 2022;75(1):25-36. doi: 10.4097/kja.21209 [published
489 Online First: 20220118]
- 490 23. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver
491 operating characteristic curve. *The Lancet Digital Health* 2022;4(12):e853-e55. doi:
492 10.1016/S2589-7500(22)00188-1
- 493 24. Nematzadeh S, Kiani F, Torkamanian-Afshar M, Aydin N. Tuning hyperparameters of
494 machine learning algorithms and deep neural networks using metaheuristics: A
495 bioinformatics study on biomedical and biological cases. *Comput Biol Chem*
496 2022;97:107619. doi: 10.1016/j.compbiolchem.2021.107619 [published Online First:
497 20211224]

- 498 25. Mohsen F, Al-Absi HRH, Yousri NA, et al. A scoping review of artificial intelligence-
499 based methods for diabetes risk prediction. *npj Digital Medicine* 2023;6(1):197. doi:
500 10.1038/s41746-023-00933-5
- 501 26. Imputation Analysis of Central Tendencies for Classification. 2021 IEEE International
502 IOT, Electronics and Mechatronics Conference (IEMTRONICS); 2021 21-24 April
503 2021.
- 504 27. Kim JC, Chung K. Multi-Modal Stacked Denoising Autoencoder for Handling Missing
505 Data in Healthcare Big Data. *IEEE Access* 2020;8:104933-43. doi:
506 10.1109/ACCESS.2020.2997255
- 507 28. Missing Values Imputation Using Fuzzy C Means Based On Correlation of Variable.
508 2020 International Conference on Computational Intelligence (ICCI); 2020 8-9 Oct.
509 2020.
- 510 29. Wassan S, Suhail B, Mubeen R, et al. Gradient Boosting for Health IoT Federated
511 Learning. *Sustainability* 2022;14(24):16842.
- 512 30. Padmanabhan S, Tran TQB, Dominiczak AF. Artificial Intelligence in Hypertension:
513 Seeing Through a Glass Darkly. *Circ Res* 2021;128(7):1100-18. doi:
514 10.1161/circresaha.121.318106 [published Online First: 20210401]
- 515 31. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk
516 prediction models is infrequent and reveals worse prognostic discrimination. *J Clin*
517 *Epidemiol* 2015;68(1):25-34. doi: 10.1016/j.jclinepi.2014.09.007 [published Online
518 First: 20141023]
- 519 32. Smith LA, Oakden-Rayner L, Bird A, et al. Machine learning and deep learning
520 predictive models for long-term prognosis in patients with chronic obstructive
521 pulmonary disease: a systematic review and meta-analysis. *Lancet Digit Health*
522 2023;5(12):e872-e81. doi: 10.1016/s2589-7500(23)00177-2
- 523 33. Silva K, Lee WK, Forbes A, et al. Use and performance of machine learning models for
524 type 2 diabetes prediction in community settings: A systematic review and meta-
525 analysis. *Int J Med Inform* 2020;143:104268. doi: 10.1016/j.ijmedinf.2020.104268
526 [published Online First: 20200907]
- 527 34. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy
528 (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. doi:
529 10.1371/journal.pmed.1001381 [published Online First: 20130205]
- 530 35. Martens FK, Kers JG, Janssens AC. External validation is only needed when prediction
531 models are worth it (Letter commenting on: J Clin Epidemiol. 2015;68:25-34). *J Clin*
532 *Epidemiol* 2016;69:249-50. doi: 10.1016/j.jclinepi.2015.01.022 [published Online
533 First: 20150203]
- 534 36. Wessler BS, Nelson J, Park JG, et al. External Validations of Cardiovascular Clinical
535 Prediction Models: A Large-Scale Review of the Literature. *Circ Cardiovasc Qual*

- 536 *Outcomes* 2021;14(8):e007858. doi: 10.1161/circoutcomes.121.007858 [published
537 Online First: 20210803]
- 538 37. Mora T, Roche D, Rodríguez-Sánchez B. Predicting the onset of diabetes-related
539 complications after a diabetes diagnosis with machine learning algorithms. *Diabetes*
540 *Research and Clinical Practice* 2023;204:110910. doi:
541 <https://doi.org/10.1016/j.diabres.2023.110910>
- 542 38. Du Y, Rafferty AR, McAuliffe FM, et al. An explainable machine learning-based clinical
543 decision support system for prediction of gestational diabetes mellitus. *Scientific*
544 *Reports* 2022;12(1):1170. doi: 10.1038/s41598-022-05112-2
- 545 39. Leila Y, Allan T. Predicting Type 2 Diabetes Complications and Personalising Patient
546 Using Artificial Intelligence Methodology. In: Anca Pantea S, ed. *Type 2 Diabetes*.
547 Rijeka: IntechOpen 2020:Ch. 8.
- 548 40. Smith LA, Oakden-Rayner L, Bird A, et al. Machine learning and deep learning
549 predictive models for long-term prognosis in patients with chronic obstructive
550 pulmonary disease: a systematic review and meta-analysis. *The Lancet Digital Health*
551 2023;5(12):e872-e81. doi: 10.1016/S2589-7500(23)00177-2
- 552

Identification of studies via databases and registers

Identification

Records identified from databases and registers*:
Pubmed/MEDLINE (n = 247)
Google Scholar +
Handsearching (n = 209)
EMBASE (n = 701)
Scopus (n = 344)
IEEE Explore (n = 812)
Wiley (n = 200)

Records removed *before screening*:
Duplicate records removed
(n = 512)

Screening

Records screened
(n = 2001)

Records excluded**
(n = 1895)

Reports sought for retrieval
(n = 106)

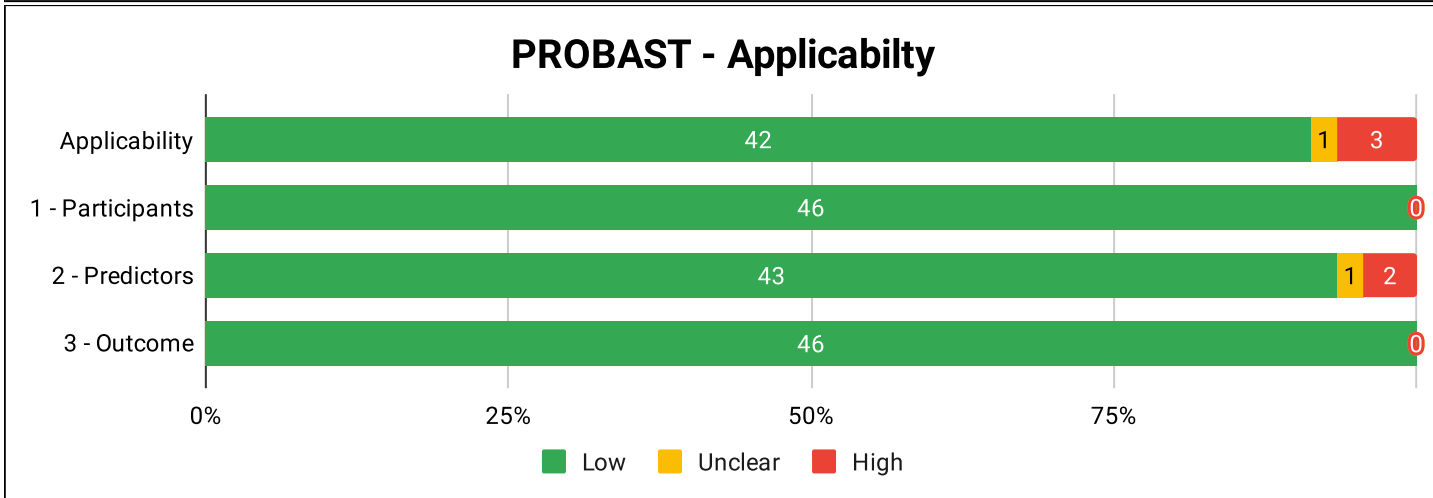
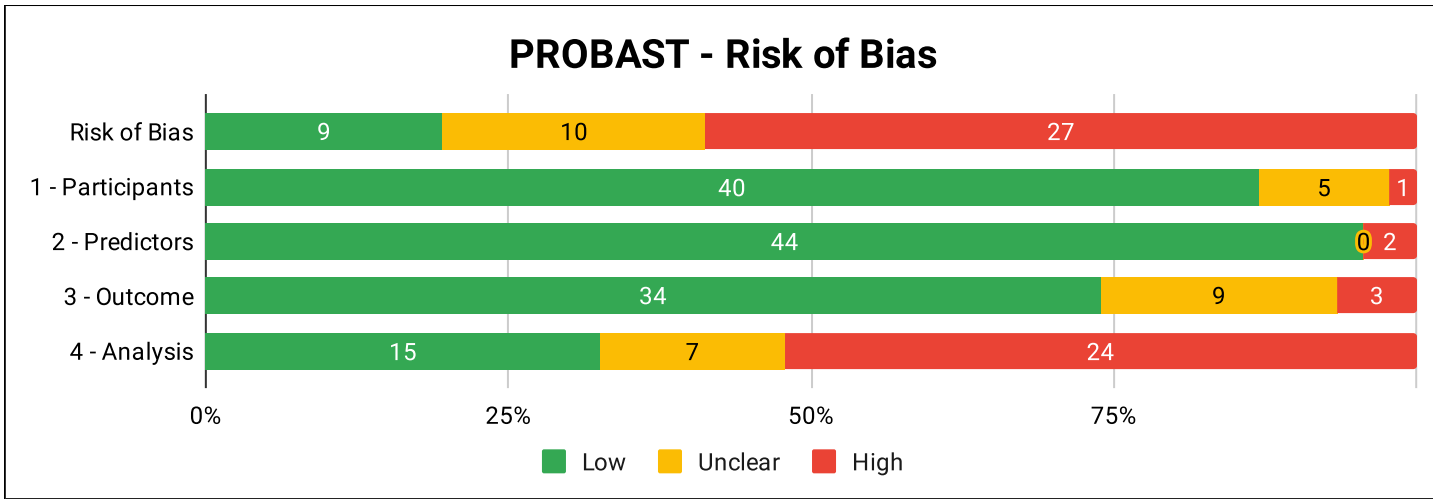
Reports not retrieved
(n = 11)

Reports assessed for eligibility
(n = 95)

Reports excluded:
n = 49
Unsuitable population : 17
Irrelevant outcome : 29
Unsuitable study design : 2
Text not in English : 1

Included

Studies included in systematic review
(n = 46)
Reports of included studies in quantitative analysis
(n = 30)



Study
 AUROC
 Standard Error

