Received: 23 June 2022 | Revised: 27 September 2022 | Accepted: 4 October 2022

DOI: 10.1111/aogs.14475

AOGS
Acta Obstetricia et Gynecologica Scandinavica

SYSTEMATIC REVIEW

Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review

Qiuyu Yang[1] | Xia Fan[2] | Xiao Cao[1] | Weijie Hao[3] | Jiale Lu[3] | Jia Wei[3] | Jinhui Tian[4,5] | Min Yin[6] | Long Ge[3,7]

[1]Evidence-Based Nursing Center, School of Nursing, Lanzhou University, Lanzhou, China

[2]Department of Obstetrics and Gynecology, The Second School of Clinical Medicine, Shanxi University of Chinese Medicine, Shanxi, China

[3]Evidence-Based Social Science Research Center, School of Public Health, Lanzhou University, Lanzhou, China

[4]Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou, China

[5]Evidence-Based Medicine Center, School of Basic Medicine Science, Lanzhou University, Lanzhou, China

[6]Health Examination Center, The First Hospital of Lanzhou University, Lanzhou, China

[7]Department of Social Medicine and Health Management, and Evidence Based Social Science Research Center, School of Public Health, Lanzhou University, Lanzhou, China

Correspondence
Min Yin, Health Examination Center, The First Hospital of Lanzhou University, 1 Donggang West Rd, Chengguan district, Lanzhou, China.
Email: ldyyym@126.com

Funding information
Gansu Provincial Science and Technology Plan Project, Grant/Award Number: 22JR5RA366 and 20CX9ZA112

Abstract

Introduction: There was limited evidence on the quality of reporting and methodological quality of prediction models using machine learning methods in preterm birth. This systematic review aimed to assess the reporting quality and risk of bias of a machine learning-based prediction model in preterm birth.

Material and methods: We conducted a systematic review, searching the PubMed, Embase, the Cochrane Library, China National Knowledge Infrastructure, China Biology Medicine disk, VIP Database, and WanFang Data from inception to September 27, 2021. Studies that developed (validated) a prediction model using machine learning methods in preterm birth were included. We used the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement and Prediction model Risk of Bias Assessment Tool (PROBAST) to evaluate the reporting quality and the risk of bias of included studies, respectively. Findings were summarized using descriptive statistics and visual plots. The protocol was registered in PROSPERO (no. CRD 42022301623).

Results: Twenty-nine studies met the inclusion criteria, with 24 development-only studies and 5 development-with-validation studies. Overall, TRIPOD adherence per study ranged from 17% to 79%, with a median adherence of 49%. The reporting of title, abstract, blinding of predictors, sample size justification, explanation of model, and model performance were mostly poor, with TRIPOD adherence ranging from 4% to 17%. For all included studies, 79% had a high overall risk of bias, and 21% had an unclear overall risk of bias. The analysis domain was most commonly rated as high risk of bias in included studies, mainly as a result of small effective sample size, selection of predictors based on univariable analysis, and lack of calibration evaluation.

Conclusions: Reporting and methodological quality of machine learning-based prediction models in preterm birth were poor. It is urgent to improve the design, conduct,

Qiuyu Yang and Xia Fan are co-first authors.

and reporting of such studies to boost the application of machine learning-based prediction models in preterm birth in clinical practice.

KEYWORDS
machine learning, prediction model, preterm birth, quality of reporting, risk of bias, systematic review

1 | INTRODUCTION

Preterm birth is usually defined as birth before 37 weeks of gestation (or 259 days).[1,2] Complications of preterm birth are the leading cause of neonatal mortality and were responsible for 35% of the world's 2.5 million deaths in 2018.[3] Preterm birth rates have increased globally, from 9.8% in 2010 to 10.6% in 2014.[4] Although most preterm babies survive, they are at increased risk of a range of long-term health problems, including chronic kidney disease,[5] hypertension,[6] diabetes,[7] ischemic heart disease,[8] and lower sleep quality.[9] Notably, preterm birth is also associated with a higher long-term risk of chronic health disorders in mothers.[10–12] Therefore, timely prevention and intervention of preterm birth is of great significance.

Machine learning is a subset of artificial intelligence that enables computer technology to learn from data to develop models and make predictions without explicit programming, which has garnered enormous attention in clinical medicine in recent years.[13,14] Machine learning techniques have been applied to predict preterm birth,[15–17] but its translation to real-world practice remains limited. One reason for this could be inadequate quality or reporting of existing studies, which led to poor transparency and reproducibility, and in turn reduced the credibility and clinical applicability. Furthermore, using a prediction model considered at high risk of bias and poor reporting might lead to unnecessary or insufficient interventions and hence affect patients' health and health systems. Accurate risk estimation of preterm birth is an essential precondition for guiding optimal management of pregnant women. Prediction models for preterm birth are a promising approach to realize risk estimation and the implementation of adequate programs for preventing preterm birth is desirable. Studies show that progesterone and pessary insertion are beneficial to prevent preterm birth in women who are at high risk for preterm birth.[18–21] However, there is limited evidence on the reporting quality and risk of bias of prediction models using machine learning methods in preterm birth. Therefore, we conducted a systematic review to assess the reporting quality and risk of bias of studies on machine learning-based prediction models in preterm birth.

2 | MATERIAL AND METHODS

Our systematic review was reported following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines.[22] This study protocol was registered with PROSPERO (CRD 42022301623).

Key message

The reporting and methodological quality of machine learning-based prediction models in preterm birth were poor.

2.1 | Search strategy and information sources

We systematically searched PubMed, Embase, the Cochrane Library, and four additional Chinese literature databases, namely, China National Knowledge Infrastructure, China Biology Medicine disk, VIP Database, and WanFang Data from inception to September 27, 2021 using a comprehensive search strategy (Table S1). Reference lists of included articles and relevant systematic reviews were also screened.

2.2 | Eligibility criteria

Publications were eligible for this systematic review based on the following inclusion criteria: (a) studies developing or validating a prediction model for preterm birth, (b) studies that typically identify as machine learning, such as artificial neural networks, decision trees, and support vector machines were included, and (c) studies using two or more predictors. The exclusion criteria were as follows: reviews, letters, or conference abstracts. We had no language restrictions.

2.3 | Data extraction

Two independent researchers (YQY, FX) screened the titles and abstracts of the identified publications. Then, we reviewed the retrieved full-text articles for eligibility. Disagreements were discussed and adjudicated by a third reviewer (LG). Three independent researchers (YQY, XC, and FX) extracted data based on the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) checklist.[23] From each included study, we extracted the following information: first author, publication year, journal, country, study objective, data source, study design, data collection period, target population, number of fetuses, number of participants and number of events, number of predictors initially considered vs that retained in the final model, definition of the outcome, modeling method, method of handling missing data, and model performance measures.

2.4 | Quality of reporting and assessment of risk of bias

The quality of reporting of the included studies was assessed independently with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.[24,25] The TRIPOD reporting guideline consists of 22 items related to the title and abstract, background and objectives, methods, results, discussion, and other information.

The risk of bias of the included studies was assessed independently by two reviewers based on the Prediction model Risk of Bias Assessment Tool (PROBAST).[26] The PROBAST included 20 signaling questions across four domains (participants, predictors, outcome, and analysis). Each signaling question was answered as yes, probably yes, no, probably no, or no information. The risk of bias judgment for each domain was rated as low, high, or unclear. The overall assessment of risk of bias was rated as low if each domain scored low, high if at least one domain was judged to be at high risk of bias, and unclear if at least one domain was judged unclear and all other domains were judged as low.

2.5 | Data synthesis

Findings were summarized using descriptive statistics and visual plots. We analyzed adherence per item and overall adherence per publication. TRIPOD adherence was calculated for each item by dividing the number of studies that adhered to a specific item by the number of studies in which the item was applicable. For each publication, TRIPOD adherence was calculated by dividing the total number of reported items by the total number of applicable reporting items. If an item was 'not applicable' for a particular study (five items were specific to external validation, including items 10c, 10e, 12, 13c, 17, and 19a; three conditional items were applicable to model development, including items 5c, 11, and 14b; two conditional items were applicable to model validation, including items 10e and 17), it was excluded when calculating the TRIPOD adherence.

3 | RESULTS

3.1 | Study selection

The search identified 10 299 articles through seven databases. A total of 8666 articles remained after duplicates were removed. A total of 8550 records were excluded after title and abstract screening and 120 underwent full-text review. Applying the selection criteria, a total of 29 articles were included. The flowchart diagram illustrating article selection is shown in Figure S1 and a list of excluded articles can be found in Table S2.

3.2 | Study characteristics

Characteristics of included studies are summarized in Table S3. Of the 29 included studies, 24 were development-only studies and five were development studies that also included an external validation of the developed model. Most studies defined preterm birth as less than 37 weeks ($n = 23$) and were developed based on retrospective cohorts ($n = 21$). Twenty studies developed models for any preterm birth, but nine focused on spontaneous preterm birth. Ten studies developed models for singleton pregnancies and five for both singleton and multiple pregnancies. It was unclear whether the models were developed for singleton and/or multiple pregnancies in 14 studies. Seven studies targeted asymptomatic pregnant women and two studies targeted symptomatic pregnant women. It was unclear whether the models were targeted for symptomatic and/or asymptomatic women in 20 studies.

Table S4 summarizes the modeling techniques and the total number of techniques reported in the included studies. One study included nine modeling techniques, one included eight modeling techniques, one included seven modeling techniques, four included six modeling techniques, one included five modeling techniques, two included four modeling techniques, five included three modeling techniques, four included two modeling techniques, and ten included one modeling technique.

3.3 | TRIPOD adherence

The completeness of items for each section of the TRIPOD statement is summarized in Table S5. Overall, six TRIPOD (sub-)items reached at least 80% adherence (rationale, design/data, setting, outcome, overall interpretation, and funding), and 14 TRIPOD (sub-) items were below 30% adherence. Figure 1 summarizes the reporting adherence for each item and Figure 2 summarizes the reporting adherence across publications. Overall, publications adhered to between 17% and 79% of the TRIPOD reporting items and had a median adherence of 49%. At least 50% TRIPOD adherence was achieved by 35% of publications overall. Scoring for each included publication is provided in Table S6.

3.3.1 | Title and abstract (items 1 and 2) and Introduction (item 3)

Two studies fully adhered to title and abstract recommendations. For included studies, the description of type of prediction model study was poorly reported but target population and outcome to be

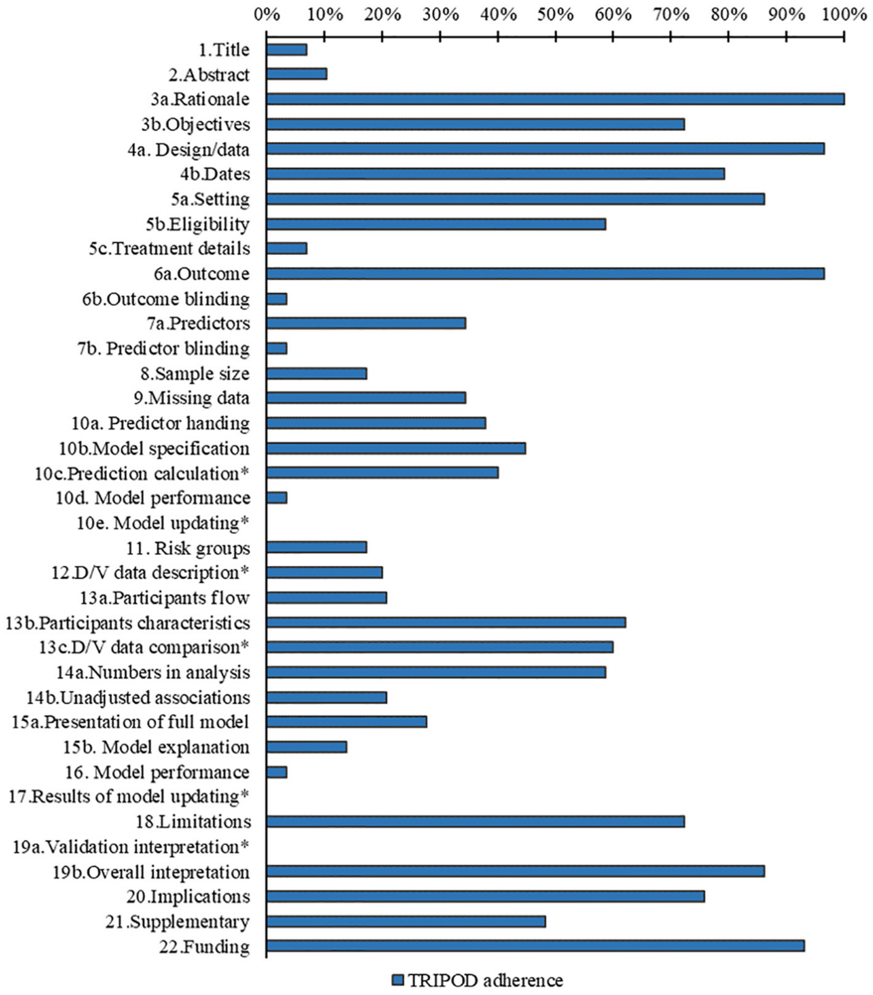

FIGURE 1 Reporting of the items of the TRIPOD statement, *item not applicable for development study

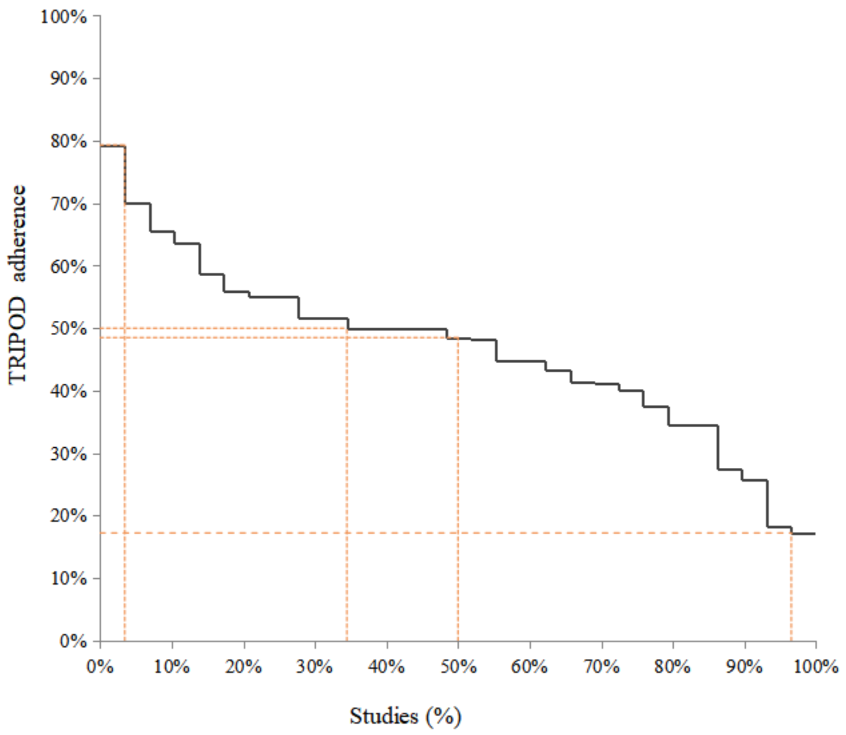

FIGURE 2 Overall adherence to TRIPOD per study

predicted were well reported in title, and the setting and predictors were poorly reported in abstract. Background was well reported for all included studies.

3.3.2 | Methods (items 4–12)

Study design and definition of outcome were well reported for all publications. The key study date was reported in 23 studies, study setting in 25 studies, model performance (calibration and discrimination) in one study, and eligibility criteria of participants in 17 studies. The definition of predictor was reported in 10 studies. Model building procedures were reported in 13 studies. Sample size justification was reported in five studies. Twenty-five studies reported the method of internal validation. Ten studies reported how missing data were handled. Five studies assessed risk groups. Blinding of outcome and predictors was reported in one study. Treatment details were reported in two studies. No study reported model updating. Differences between development and validation set such as data sources, eligibility criteria, and measurement of predictors were reported in two studies.

3.3.3 | Results (items 13–17)

Only one study fully adhered to the model predictive performance recommendations.

The numbers of participants and events were described in 17 studies. Characteristics of participants were reported in 18 studies. Eight studies presented the full model. Six studies reported the flow of participants and unadjusted associations. Three studies compared development and validation set. Four studies explain how to use the model.

3.3.4 | Discussion (items 18–20) and other information (items 21 and 22)

Overall interpretation of results was reported in 25 studies. Twenty-one studies reported limitations. Potential clinical use of the model and implications for future research were reported in 22 studies. Two development-with-validation studies compared validation performance with reference to performance in the development data. Availability of supplementary resources was mentioned in 14 studies. Funding information was reported in 27 studies.

3.4 | Risk of bias assessment

Overall assessment of risk of bias according to PROBAST was shown in Figure 3. For all included studies, 79% had a high overall risk of bias, and 21% had an unclear overall risk of bias. Risk of bias assessment for each included study is provided in Table S7.

In the participant domain, 17 studies were judged at low risk of bias, and 11 studies were at unclear risk of bias because no information was provided on the inclusions and exclusions of participants. The risk of bias was high in one study because the data were collected through cross-sectional survey. In the predictor domain, 20 studies had low risk of bias, and 7 studies were at unclear risk of bias because there was no information on whether predictor assessment was blinded to outcome data and predictors were not reported in these retrospective studies. Two studies were at high risk of bias, because postpartum information was used for model development. The outcome domain was rated as low risk of bias for all included studies. In the analysis domain, six studies were at unclear risk of bias because details on methods to handle continuous predictors and definition of categorical predictor groups, handing of missing data, accounting for complexities in the data, model overfitting, and optimism in model performance were infrequently provided. Twenty-three studies had high risk of bias because of small effective sample size (defined as events per predictor less than 10), selection of predictors based on univariable analysis, and lack of calibration evaluation.

4 | DISCUSSION

In this systematic review, we assessed the reporting quality and risk of bias of studies describing the development and validation of

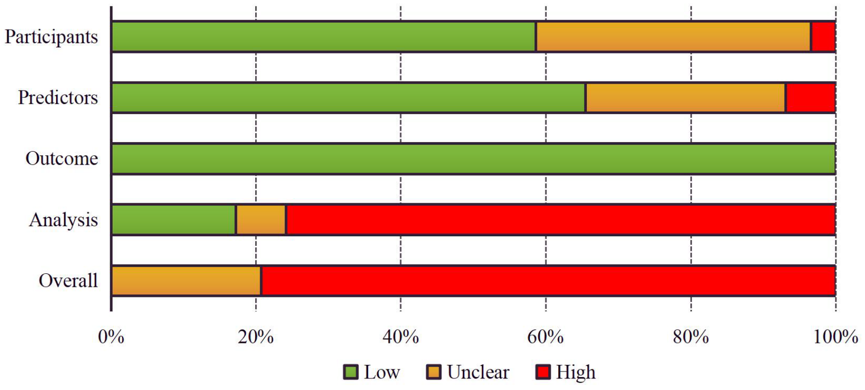

FIGURE 3 Overall risk of bias assessment of included studies using PROBAST

machine learning prediction models in preterm birth. All included studies were found to have inadequate reporting of essential items for prediction modeling. Reporting was mostly poor in some sections including title and abstract, blinding of outcome and predictors, sample size justification, differences between development and validation set, flow of participants, presentation of full model, explanation of model, and model performance. Overall, 79% had a high overall risk of bias, and 21% had an unclear overall risk of bias. The analysis domain was most commonly rated as high risk of bias in the included studies, mainly as a result of small effective sample size, selection of predictors based on univariable analysis, and lack of calibration evaluation.

Overall, most studies failed to be fully compliant with the title and abstract reporting recommendations in TRIPOD, which can affect the identification and screening of articles by potential readers or systematic reviewers. The information of blinding of outcome and predictors was rarely reported in studies. Given that preterm birth is a clearly defined outcome, the focus for prediction models in preterm birth was assessed predictors without knowledge of the outcome. Therefore, the information of blinding of predictors should be explicitly stated in future studies. The limited effective sample sizes increased the risk of overfitting and likely led to overoptimistic prediction estimations for many models. In addition, differences between development and validation set, such as data sources, eligibility criteria, and measurement of predictors, were poorly reported in studies, which might potentially affect model transportability.[24] Few studies presented the full model and explained how to use the model. This could affect subsequent validation studies or clinical practice.

Most studies had high risk of bias in the analysis domain for small effective sample size. Small effective sample sizes were likely to lead to overfitting and underfitting of the model, which were likely to yield biased estimates of the model's apparent predictive performance.[26] In addition, selection of predictors based on univariable analysis was one cause of high risk of bias because the method could lead to incorrect predictor selection. Furthermore, inappropriate evaluation of relevant model performance (calibration and discrimination) would also lead to high risk of bias because the ability or performance of the model to provide accurate individual probabilities was not completely known.[26]

To our knowledge, there was limited evidence on the reporting quality and risk of bias of published studies for the development or validation of machine learning prediction models in preterm birth. In contrast to prediction models in other medical areas, title and abstract, sample size justification, description of flow and baseline characteristics of participants, and the presentation of the full models were poorly reported in prediction models based on machine learning methods in oncology.[27] A systematic review assessed reporting quality of prediction models using supervised machine learning in many medical areas such oncology, neurology, and surgery.[28] This review found that some essential items were inadequately reported in publications, especially in title and abstract, blinding, model building procedures, model specifications, and model performance. The results of these studies were partially similar to our findings, with poor reporting in title and abstract section. Furthermore, we observed that differences between development and validation and explanation of model were also infrequently reported. A systematic review of 152 studies found that the analysis domain was most commonly rated as high risk of bias in prediction models.[29] This review found deficiencies in number of participants with the outcome, overfitting, and handling of participants with missing data. Another publication about machine learning-based prediction models in living organ transplantation also considered 39/49 studies as at high risk of bias because of small sample size, mishandling of missing data, weak strategies for model building, and model performance evaluation.[30] Similar to our results, major deficiencies were found in the analysis domain including number of participants with the outcome, selection of predictors, and evaluation of relevant model performance.

The systematic review highlighted the common reporting and methodological flaws found in machine learning-based prediction models in preterm birth. We used systematic methods that included a robust search strategy of seven databases with independent study selection and extraction by two researchers. Although a systematic search for studies, it was possible that some eligible articles might have been missed for inadequate reporting or inconsistent terminology of prediction models in title and abstract. However, our aim for this review was to describe the reporting and methodological flaws of current prediction models in preterm birth. Given that our findings are comparable to the reviews above,[27–30] it is unlikely that additional studies would change our conclusion. We used the TRIPOD statement, which was designed for regression-based prediction models, to evaluate the reporting quality of prediction models developed using machine learning. Though some items of this statement may be difficult to adhere to, such as presentation of the full model, the vast majority of the items in the statement were relevant to machine learning-based prediction model studies. Similarly, the PROBAST tool was not fully applicable to machine learning-based models for assessing the risk of bias because of some different approaches to development and validation, and terminology.

Complete reporting allowed studies to be understood, replicated, and used. However, prediction models based on machine learning were much more reliant on computers for implementation of the underlying model, which was often labeled as a black box, and led to poor transparency and poor reproducibility. Therefore, the TRIPOD collaboration has initiated the development of a TRIPOD Statement and PROBAST quality assessment tool specific to machine learning (TRIPOD-AI and PROBAST-AI).[31,32] Periodic reviews and re-reviews of prediction models are warranted in preterm birth to continue to assess the quality of reporting and methodology as the rapid and constant evolution of machine learning continues. Furthermore, sample size contributed largely to the high risk of bias, so future methodological research could focus on how to calculate appropriate sample sizes for the machine learning technique.

5 | CONCLUSION

Reporting and methodological quality of machine learning-based prediction models in preterm birth were poor. Guidance for machine learning-based prediction models is urgently needed. Particular areas for which reporting needs to be improved include the title and abstract, blinding of predictors, sample size justification, explanation of model, and model performance. Factors contributing to risk of bias include small effective sample size, selection of predictors based on univariable analysis, and lack of calibration evaluation.

AUTHOR CONTRIBUTIONS

GL, YM, and TJH conceived and designed the study. TJH, YQY, and FX developed the search strategy, and carried out the screening. YQY, CX, and FX extracted the data of all items from all articles. YQY, FX, HWJ, LJL, and WJ performed the analysis. YQY and FX produced the first draft. GL and YM critically reviewed and edited the article. All authors read and approved the final manuscript.

FUNDING INFORMATION

Gansu Provincial Science and Technology Plan Project (22JR5RA366, 20CX9ZA112).

CONFLICT OF INTEREST

The authors have stated explicitly that there are no conflicts of interest in connection with this article.

ORCID

Min Yin https://orcid.org/0000-0002-2435-6111
Long Ge https://orcid.org/0000-0002-3555-1107

REFERENCES

1. WHO: recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. Modifications recommended by FIGO as amended October 14, 1976. *Acta Obstet Gynecol Scand*. 1977;56:247-253.
2. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. *Lancet*. 2008;371:75-84.
3. United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME). *Levels and Trends in Child Mortality: Report 2019, Estimates Developed by the United Nations Inter-Agency Group for Child Mortality Estimation*. United Nations Children's Fund; 2019.
4. Chawanpaiboon S, Vogel JP, Moller AB, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health*. 2019;7:e37-e46.
5. Crump C, Sundquist J, Winkleby MA, Sundquist K. Preterm birth and risk of chronic kidney disease from childhood into mid-adulthood: National Cohort Study. *BMJ*. 2019;365:l1346.
6. Crump C, Sundquist J, Sundquist K. Risk of hypertension into adulthood in persons born prematurely: a national cohort study. *Eur Heart J*. 2020;41:1542-1550.
7. Crump C, Sundquist J, Sundquist K. Preterm birth and risk of type 1 and type 2 diabetes: a national cohort study. *Diabetologia*. 2020;63:508-518.
8. Crump C, Howell EA, Stroustrup A, McLaughlin MA, Sundquist J, Sundquist K. Association of preterm birth with risk of ischemic heart disease in adulthood. *JAMA Pediatr*. 2019;173:736-743.
9. Visser SSM, van Diemen WJM, Kervezee L, et al. The relationship between preterm birth and sleep in children at school age: a systematic review. *Sleep Med Rev*. 2021;57:101447.
10. Crump C, Sundquist J, McLaughlin MA, Dolan SM, Sieh W, Sundquist K. Pre-term delivery and long-term risk of heart failure in women: a national cohort and co-sibling study. *Eur Heart J*. 2021;43:895-904.
11. Crump C, Sundquist J, Sundquist K. Preterm delivery and long term mortality in women: National Cohort and Co-Sibling Study. *BMJ*. 2020;370:m2533.
12. Crump C, Sundquist J, Sundquist K. Preterm delivery and long-term risk of stroke in women: a National Cohort and Cosibling study. *Circulation*. 2021;143:2032-2044.
13. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019;188:2222-2239.
14. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216-1219.
15. Bahado-Singh RO, Sonek J, McKenna D, et al. Artificial intelligence and amniotic fluid multiomics: prediction of perinatal outcome in asymptomatic women with short cervix. *Ultrasound Obstet Gynecol*. 2019;54:110-118.
16. Lee KS, Ahn KH. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. *J Korean Med Sci*. 2019;34:e128.
17. Arabi Belaghi R, Beyene J, McDonald SD. Prediction of preterm birth in nulliparous women using logistic regression and machine learning. *PLoS One*. 2021;16:e0252025.
18. Care A, Nevitt SJ, Medley N, et al. Interventions to prevent spontaneous preterm birth in women with singleton pregnancy who are at high risk: systematic review and network meta-analysis. *BMJ*. 2022;376:e064547.
19. Conde-Agudelo A, Romero R, Da Fonseca E, et al. Vaginal progesterone is as effective as cervical cerclage to prevent preterm birth in women with a singleton gestation, previous spontaneous preterm birth, and a short cervix: updated indirect comparison meta-analysis. *Am J Obstet Gynecol*. 2018;219:10-25.
20. EPPPIC Group. Evaluating progestogens for preventing preterm birth international collaborative (EPPPIC): meta-analysis of individual participant data from randomised controlled trials. *Lancet*. 2021;397:1183-1194.
21. Rahman RA, Atan IK, Ali A, et al. Use of the Arabin pessary in women at high risk for preterm birth: long-term experience at a single tertiary center in Malaysia. *BMC Pregnancy Childbirth*. 2021;21:368.
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
23. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11:e1001744.
24. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162:W1-W73.
25. Collins GS, Reitsma JB, Altman DG, Moons KGM, members of the TRIPOD group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol*. 2015;67:1142-1151.
26. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170:W1-W33.

27. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol*. 2021;138:60-72.
28. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22:12.
29. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281.
30. Haller MC, Aschauer C, Wallisch C, et al. Prediction models for living organ transplantation are poorly developed, reported, and validated: a systematic review. *J Clin Epidemiol*. 2022;145:126-135.
31. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577-1579.
32. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11:e048008.

SUPPORTING INFORMATION

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

How to cite this article: Yang Q, Fan X, Cao X, et al. Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review. *Acta Obstet Gynecol Scand*. 2023;102:7-14. doi:10.1111/aogs.14475