The efficacy of machine learning in predicting fetal growth restriction: a systematic review and meta-analysis

Chenhan Zheng  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Ji Chunya  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Lin Guimei  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Yin Linliang (yllsznthello@hotmail.com)  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Deng Xuedong  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Wang Benjing  
The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital

Research Article

Keywords: fetal growth restriction, machine learning, predictive model, pregnancy, predictor

Posted Date: December 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2307066/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Abstract

Objectives

Fetal growth restriction (FGR) is one of the most common causes of perinatal death and various short-and long-term complications. Accurate identification of fetal growth restriction is essential to reduce adverse perinatal outcomes. With the development of machine learning, many predictive models for fetal growth restriction have emerged. We assessed their performance by conducting systematic reviews and meta-analyses.

Methods

A systematic literature search was performed for relevant studies reported before May 31, 2022. The quality of the studies was assessed in strict accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (version 2020). Models with the reported area under the receiver operating characteristic (AUROC) indicator curve were meta-analyzed to determine the factors that contributed most to model performance.

Results

After the screening, 20 papers were eligible for synthesis, 19 were rated as high risk of bias, and 1 was rated as unclear risk of bias. From which 37 models were extracted, the c-statistic of the pooled random effects model \((I^2 = 96.0\%)\) was 0.8123 (0.7824–0.8433, 95%CI). Among them, the c-statistic of the pooled risk model using the consensus developed by the Delphi procedure as the defined model was 0.81 (0.77–0.85, 95%CI), and the c-statistic of the pooled risk model using other defined models was 0.83 (0.78–0.87, 95%CI). In selecting predictors, most models were constructed by combining basic maternal characteristics, maternal pregnancy radionics, and serological indicators during pregnancy.

Conclusion

Machine learning has an excellent predictive value for FGR, which indicates that practical machine learning can be used as a possible means of FGR identification. However, improvements are needed in terms of quality and study design. We look forward to a unified definition and establishing an effective scoring tool to identify FGR accurately and intervene in time.

Objectives

Fetal growth restriction (FGR) is affected by maternal, fetal, placental, and other pathological factors, and fetal growth does not reach its genetic potential. Fetuses with FGR are at higher risk for perinatal
morbidity and mortality. They may have long-term adverse outcomes, such as cognitive impairment in childhood and increased risk of cardiovascular and endocrine diseases in adulthood. Fetal growth restriction is one of obstetricians' most common pregnancy complications, affecting approximately 3-9% of pregnancies. The incidence of FGR is different in different countries, which is about 8.77% in China. Studies have shown that prenatal identification of FGR reduces mortality and morbidity. A study involving more than 92,000 single babies found that the stillbirth rate was lower with prenatally detected FGR (9.7%) than undetected FGR.

Guidelines issued by the American College of Obstetricians and Gynecologists (ACOG) defined fetuses with the estimated fetal weight (EFW) < 10th percentile of corresponding gestational age as FGR. While guidelines from the Royal College of Obstetricians and Gynecologists (RCOG) and the Canadian College of Obstetricians and Gynecologists (SCOG) explain FGR as affected by medical factors (maternal factor, fetal factor, placental diseases, etc.), fetal growth not reaching its genetic potential, EFW or abdominal circumference (AC) below the 10th percentile for gestational age. Gordijn et al. developed a consensus definition of FGR using Delphi Procedure (starting now referred to as DP), which defines early onset and late onset, respectively. For late-onset FGR, if EFW or AC is less than the third percentile, it can be directly defined. In addition, two or more of the following three items: AC/EFW < 10th centile, AC/EFW crossing centiles > 2 quartiles on growth centiles, CPR < 5th centile or UA-PI > 95th centile can also be defined as FGR. The international definition of FGR has no unified gold standard now often, EFW by ultrasound or AC is lower than the 10th percentile of weight for the corresponding gestational age is often defined, which coincides with small-for-gestational-age fetuses SGA. However, most SGA exist only have small growth parameters, they have low risk of adverse pregnancy outcomes. However, in reality, for fetuses whose growth parameters are lower than the 10th percentile, perinatal monitoring and intervention are generally needed to reduce the incidence of adverse pregnancy outcomes. Therefore, accurate prenatal identification of FGR plays an essential role in reducing adverse perinatal outcomes and economic costs.

In recent years, the development of machine learning has been booming, used in the large-scale, complex algorithm to run on different kinds of data sets to identify and summarize meaningful patterns. It has been widely used in various fields, including medicine, for the future more accurate early diagnosis, personalized therapy provides a possibility to improve the quality of the population has an important significance. The adverse perinatal outcomes caused by FGR and its impact on long-term prognosis have received widespread attention in recent years. Some researchers have applied machine learning to predict fetal growth restriction, but the prediction results and predictors are highly heterogeneous. There is still a lack of empirical evidence to explore the predictive value of machine learning for FGR. Therefore, we conducted this systematic review and meta-analysis to evaluate the application value.

**Methods**

Our Systematic review and meta-analysis were carried out in strict accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Guidelines (2020 version) and
prospectively registered at PROSPER (ID: CRD42022335779).

**Retrieval Strategy**

We conducted a comprehensive and systematic search of PubMed, Embase, Cochrane Library, and Web of Science. The search date was up to May 31, 2022, and there were no restrictions on regions and languages. The search method was subject words plus free words, with the detailed search strategy as follows: ( (Fetal Growth Retardation) OR (Intrauterine Growth Retardation) OR (Growth Retardation, Intrauterine) OR (Intrauterine Growth Restriction) OR (fetal growth abnormalities)) AND ((machine learning) OR (prediction) OR (risk model) OR (screening) OR (algorithms) OR (random forest) OR (ANN) OR (SVM) OR (nomogram))

**Inclusion And Exclusion Criteria**

This systematic review established the following inclusion criteria :(1) Fetal intrauterine growth restriction; (2) A complete prediction model is constructed; (3) There are indicators to evaluate the performance of the prediction model, such as c-index, sensitivity, specificity, etc.

At the same time, we formulated the following exclusion criteria :(1) Only risk factors were analyzed in the study, but the prediction model was not constructed; (2) The prediction model was constructed in the study, but its prediction efficiency was not evaluated, or AUC, sensitivity, and specificity were missing; (3) Case reports, letters, conference abstracts, systematic reviews or meta-analyses, consensus statements, guidelines, and reviews, etc.

**Literature screening and data extraction**

In the data extraction process, the retrieved literature is first imported into ENDNOTE software, the duplicate literature is eliminated, and the original studies that meet the preliminary requirements are screened according to the title and abstract. The two researchers then went on to manually search the references of each eligible article for potential other eligible studies. Finally, the two researchers cross-checked and came up with the final included literature. In the process of analyzing data and extracting information, we develop primary information extraction forms, including title, author, a fixed number of years of the publication, the authors state, research types, sample source, positive sample - training set, the training set of sample, external validation, positive sample - validation set, validation set sample, forecast factor selection method, model type, model number, model variables, outcome indicators, and other.

The above literature screening and information extraction were carried out by two researchers (Zheng and Ji) independently and cross-checked after completion. In case of any dispute, the third researcher was asked to assist in adjudicating.
Quality Evaluation

PROBAST\textsuperscript{10} was used to evaluate the risk of bias in the included original studies, which included a large number of questions in four fields: participants, predictor variables, results, and statistical analysis, reflecting the overall risk of bias and overall use. The four domains contained 2, 3, 6, and 9 specific questions, respectively, with three responses to each question (yes/probably yes, no/probably no, and no information). A domain is considered at high risk if it contains at least one question indicating no or yes. All yes or probably yes questions should be included to be regarded as low risk. The overall risk of bias was rated as low risk when all areas were considered low risk and high risk when at least one area was considered high risk.

Two investigators (Zheng and Ji) independently assessed the risk of bias based on PROBAST and cross-checked at the end. If there was a dispute, a third investigator (Yin) was asked to assist in adjudicating.

Data Synthesis And Statistical Methods

We conducted a meta-analysis of the indicators (c-index and accuracy) for model evaluation of machine learning models. If the c-index was missing 95% confidence intervals and standard errors, we estimated the standard errors according to the study of Debray TP et al. (Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019;28:2768-86.). In case of lack of accuracy in the original study, we will calculate based on sensitivity and specificity, combined with the number of samples of each molecular subtype and the number of modeling samples. Given the differences in the variables included and the inconsistent parameters among the machine learning models, we prioritized using the random effects model in the meta-analysis. Our meta-analysis of this study was implemented in R4.2.0 (R Development Core Team, Vienna, http://www.R-project.org).

Results

Literature retrieval

A total of 15649 literature were retrieved, and the duplicated literature and some irrelevant literature (including guidelines, reviews, case reports, in vitro experiments, conference abstracts, letters, etc.) screened by automation were removed, leaving 6295. After screening the initially qualified original studies according to their titles and abstracts, the full text of 51 literature was read, and the final 20 literature met the inclusion criteria (Fig. 1). A total of 37 models were included in the 20 papers. For the same definition of FGR model, only the model with the best prediction performance described by the author of the original paper was included in this study. In addition, if the models for prediction in the first trimester, the second trimester, and the third trimester were described in the study, we included the models with the best performance in different periods. In all the papers, AUROC is reported as its performance index.
Characteristics Of Included Literature

Among the 20\textsuperscript{11–30} articles that were finally eligible (Table 1), 4\textsuperscript{11,13,18,29} were from multicenter studies, and the other 16 were from single-center studies. Four \textsuperscript{11,13,19,29} studies were externally validated, and two \textsuperscript{11,29} were also internally validated at the same time; the rest were not validated.

Among all the literature, one original study \textsuperscript{25} did not specify the definition of FGR in the study. Ten studies \textsuperscript{11,12,16,18,19,22–24,27,29} which used DP defined FGR as birth weight (BW)/EFW less than the 10th percentile of normal fetus and associated with prenatal uteroplacental insufficiency (umbilical artery (UA) PI, uterine artery (UtA) PI, or cerebroplacental ratio (CPR) abnormalities). Nine \textsuperscript{11,12,16,18,19,22–24,29} included BW/EFW less than the third percentile of normal as inclusion criteria. Nine studies \textsuperscript{13–15,17,20,21,26,28,30} only used weight as the definition, and five of them \textsuperscript{14,15,17,20,30} were defined as BW/EFW less than the 10th percentile of normal weight. Four papers \textsuperscript{13,21,26,27} defined FGR as BW/EFW less than the third or fifth percentile of normal.

In the selection of the study population, the target population of 1 article \textsuperscript{12} was pregnant women with abnormal Down’s screening index in the first trimester, 1 article \textsuperscript{21} was from pregnant women with early-onset preeclampsia, and 1 article \textsuperscript{28} was from pregnant women with risk factors of hypertension. In the prediction of outcome indicators, one paper \textsuperscript{12} predicted early FGR, four papers \textsuperscript{16,18,19,24} predicted late FGR, one paper \textsuperscript{23} predicted early FGR and late FGR, respectively, and the rest did not distinguish between early and late cases.

For the selection of predictors and the establishment of models, seven papers \textsuperscript{14,16–18,22,24,28} chose the reverse stepwise multiple logistic regression to determine the final model, and three papers \textsuperscript{13,23,30} chose the forward stepwise logistic regression to determine the final model, and the rest did not clearly explain.

Characteristics Of Predictors

Except for the unreported specific predictors of the six deep learning models, the features of the remaining models can be found in the original paper and supplementary materials. To avoid repeated use of model variables by multiple models proposed in the same study, which may lead to the high frequency of predictors, we only included predictors used in the same study once. The model used in the number of features from 2 to 14, 22 model forecast variables contain ultrasonic indicators, three models to predict variable contains maternal gene expression level, 16 models predict variable contains maternal pregnancy complications, Body Mass Index (BMI), history of FGR, pregnancy bad living habits such as smoking status and basic characteristics, 12 models predict variable contains maternal serological indexes. The most common predictors of maternal characteristics were height, BMI, mean arterial pressure in the first trimester, smoking history, hypertension, diabetes, autoimmune disease, previous adverse pregnancy history, and race; The most common predictors of radiomics were: UtA RI in the second trimester, UA RI in the second trimester and UTA-PI in the first trimester; For maternal serological
predictors, the most common were placental growth factor (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1)/PIGF ratio, and pregnancy-associated plasma protein-A (PAPP) (Table 2).

**Quality Assessment**

The risk of bias was evaluated according to PROBAST, and the specific results of the risk of bias were shown in the Table S1. One of the 20 papers was rated as having an ambiguous risk of bias, and the other papers were rated as having a high risk of bias. In selecting research objects, 6 papers used retrospective studies and had a high risk of bias, and three included pregnant women with high-risk factors that we judged to be at high risk of bias. All studies were considered to have a low risk of bias in the risk assessment of predictors and results. Nearly all results used the more accepted definition of EFW below the 10th percentile, but because there is no gold standard in the diagnosis of FGR, the definition of diversity may increase the risk of deviation of the model; as a result, all the papers have an unclear risk of bias about the reference standard. In terms of the risk of bias in statistical analysis, all the studies have a high risk of bias, most of the prediction models of the studies lack verification support, and some models also have the problem that the positive number is less than ten times the number of variables.

**Model Comprehensive Detection Performance**

Finally, 37 models in 20 studies were included in the meta-analysis. For FGR cases, the C-statistic of the training set’s random effects model ($I^2 = 96.0\%$) was $0.8123$ (0.7824–0.8433, 95%CI), and its range was $0.620–0.974$. The sensitivity was $0.77$ (0.71–0.81, 95%CI), $I^2 = 89.9\%$, and the specificity was $0.82$ (0.78–0.86, 95%CI), $I^2 = 98.6\%$. The C-statistic of the random effects model summarized by the validation set ($I^2 = 84.6\%$) was $0.7666$ (0.7235–0.8124, 95%CI), and its range was $0.620–0.920$. The sensitivity of the validation set of the training set was $0.75$ (0.70–0.80, 95%CI). $I^2 = 0\%$, specificity was $0.82$ (0.80–0.84, 95%CI), $I^2 = 36.0\%$. For FGR cases defined by DP, the C-statistic of the pooled risk model was $0.81$ (0.77–0.85, 95%CI), $I^2 = 95\%$, its sensitivity was $0.77$ (0.74–0.80), $I^2 = 44.68\%$, and its specificity was $0.81$ (0.77–0.84). $I^2 = 98.24\%$. For FGR cases defined by other methods, the C-statistic of the pooled risk model was $0.83$ (0.78–0.87, 95%-CI), $I^2 = 93\%$, and its sensitivity was $0.72$ (0.63–0.80), $I^2 = 92.17\%$, and specificity was $0.83$ (0.75–0.88). $I^2 = 98.91\%$. (Fig. 2)

**Discussion**

This is a systematic review of research on the effects of machine learning on predicting FGR. A total of 20 literature and 37 models were included, and almost every model showed good predictive performance. The predictors with predictive value in these models proved to have certain clinical value in many clinical studies and were recognized in clinical practice.
Heterogeneity Defined By FGR

In the literature included in this study, the definitions of FGR varied, but all used EFW, birth weight, or growth parameters as the standard. At present, the clinical identification of FGR mainly relies on fetal AC and EFW, which may lead to overdiagnosis of FGR and unnecessary clinical intervention. Some of these fetuses have already played their due growth potential, and their small size is mainly due to genetic factors from their parents. Gordijn et al. 7 in 2015 developed a consensus definition of FGR using Delphi Procedure introduced abnormal growth curves and the indicator of uteroplacental insuiciency. The management recommendations on fetal growth restriction issued by International Federation of Gynecology and Obstetrics(FIGO) 31 in 2021 confirmed the value of this definition. Such a definition is valuable because fetuses with FGR exclude chromosomal abnormalities, most of which are caused by uteroplacental dysfunction 1,32. Adding evidence of uteroplacental dysfunction can better identify fetuses at risk of adverse pregnancy outcomes and prevent some fetuses with EFW < 10% from being overdiagnosed, true fetal growth retardation would be underestimated. Because FGR affects perinatal outcomes and produces long-term complications, accurate identification of FGR and prediction of FGR patients are of great significance for developing individualized monitoring and intervention measures. In this study, most studies referenced this consensus to more accurately identify fetuses that are not reaching their growth potential and are at risk for adverse outcomes. The machine learning models presented in these studies provide a more rigorous definition approach to identify FGR with superior individual performance.

Heterogeneity And Aggregation Performance

Our results based on 37 models show that machine learning has good predictive value in FGR, mainly reflected in AUC, sensitivity, and specificity. However, there is a considerable risk of bias in the modeling process, and the reasons are generally as follows: (1) Insufficient models for verification: Although FGR has a relatively high incidence in clinical practice, most of the included literature in this study are single-center studies, which are insufficient to support external verification. External validation for predicting model of clinical effectiveness evaluation is critical, not validated model mostly in clinical performance is poorer, because the center of the other general comparison of prediction model used to develop the overall due to factors such as geographical, clinician decisions will exist certain differences, so the forecasting model is more support for external validation; (2) There are differences in the definition of FGR: there is no unified international standard for FGR at present, so multiple definitions can be adopted, which may be the main reason for the heterogeneity. Ideally, the overall machine learning performance would be calculated by pooling the AUC of all models. However, different definitions of FGR in different studies will result in significant heterogeneity to a certain extent; (3) The ratio of the number of predictors to positive events: Ideally, the prediction model should be calculated based on many positive events. And we are included in most of the models, a variable number of events is less than the normally recommended values of each forecaster's ten events. However, this is just a rule of thumb rather than the absolute standard. Still, each forecast model of low event rate may deviate from the correlation
coefficient of the model, especially for not based on the assumption of biological mechanisms of previous data or trusted connection priority. In addition, through the assessment of bias risk by PROBAST, it was also observed that there was heterogeneity in some studies on pregnant women included in the study, data preprocessing, and other aspects.

Concerning the selection of predictors and the building of models, most studies follow the recommended approach. Although there is no accepted best variable selection method, either the "full model" approach (including all candidate predictors) or the backward elimination strategy seems to be the most reliable approach. Forward selection and univariate predictor-outcome association selection (selected by significance tests) were more likely to introduce bias and overfitting but were employed in a small number of models.

Model Analysis And Predictor Analysis

This study included 20 literature and 37 models, and almost every model showed good predictive performance. Most of these models combined maternal risk factors, prenatal ultrasound detection, and biochemical indicators to predict FGR. Therefore, it is necessary to construct a risk model for predicting FGR, which should include maternal radionics, maternal basic characteristics, and serological indicators. Most of the literature compared models that simply included maternal risk factors and models that combined multiple predictors. The latter showed significant improvement in AUC, sensitivity, and specificity.

Most of the predictive covariates in our meta-analysis, such as maternal risk factors, biophysical measures, or biochemical measures, were considered to be associated with FGR. The most commonly used biochemical indicators in this study were PI GF, sFlt-1 /PI GF ratio, and PAPP. As FGR produces severe pathology of placental insufficiency, uteroplacental insufficiency is caused by hypoxia-reperfusion injury of placental villi caused by decreased or unstable uteroplacental blood flow. Placental villi are destroyed during normal development, impinges the secretion of pro-angiogenic placental growth factor (PI GF). It also promotes the secretion of soluble FMS-like tyrosine kinase-1 (sFlt-1), an anti-angiogenic protein, so these pathological changes in the placenta can be identified by changes in PI GF and SFLT-1 in maternal blood. It is worth mentioning that in some models these two indicators are measured in the first trimester, which further indicates that the underlying pathophysiology begins in the first trimester. In addition, some aneuploidy screening indicators during early pregnancy, such as pregnancy-associated plasma protein A(PAPP-A), have also been associated with placental insufficiency. For biophysical indicators, Doppler parameters such as uterine artery and umbilical artery Doppler are most commonly used in the model included in this study. Abnormalities of these parameters have a specific value in identifying placental insufficiency in most clinical studies. With the maturity of three-dimensional ultrasound technology, it is possible to measure placental volume. Although the relationship between placental volume and FGR in the first trimester is controversial, the model in the study of Papastefanou I et al. has further improved its prediction performance due to the inclusion of placental volume in the
first trimester. In two other studies \(^{27,30}\), MRI functional imaging of the placenta in the second and third trimesters also showed good predictive performance.

In the study of L. Ormesher et al. \(^{12}\), the study subjects were pregnant women with abnormal maternal serum biomarkers (low PAPP-A and elevated \(\beta\)-HCG /Inhibin/\(\alpha\)FP) during aneuploidy screening in early pregnancy and combined with the results of ultrasound examination at 21–24 weeks; it was found to have an excellent predictive value for early-onset FGR. However, the prediction efficiency for late-onset FGR is not good. In previous studies \(^{35}\), some scholars believed that women with abnormal maternal serum biomarkers have an increased risk of placental diseases. In the other two studies, the subjects were hypertensive pregnant women \(^{28}\) with early-onset preeclampsia \(^{21}\). The decreased or unstable uteroplacental blood flow in hypertensive pregnant women can lead to hypoxia-reperfusion injury of placental villi, which usually also causes preeclampsia \(^{36}\), which are all risk factors for FGR.

Unexpectedly, two papers \(^{11,29}\) were found to predict FGR by detecting biomarkers of maternal cfDNA fragments, which were found to have good predictive performance after a series of validation. Chan et al. found that the end position of cfDNA fragment in maternal plasma was related to the source of the placenta, and FGR was caused mainly by placental insufficiency. Therefore, NIPT data may provide useful information and have specific feasibility. It is easier to analyze biomarkers of cfDNA fragments without considering economic factors directly. Still, there are few relevant studies, and we look forward to verification in future multicenter studies. In addition, Ulla et al.'s original study \(^{13}\) conducted a non-targeted metabolomics study on maternal serum, screening four biological metabolites. Their ratio was used to predict FGR, and the prediction performance was higher than that of the traditional sFlt1: PlGF, but larger samples were needed to verify its efficacy.

**Strengths And Limitations**

The advantage of this study is that this is the first system to list the FGR prediction model of research, combined with clinical and more technical work and performance from the aspect of clinical evaluation, explore the application of machine learning in FGR at the same time analysis the predictors of the frequency of the various prediction models for the future study of FGR prediction model design and update offer some guidance. In addition, we searched four major databases, and any missed studies were unlikely to change our main findings significantly.

The limitations of this study are as follows : (1) even though a comprehensive database search has been conducted, machine learning has only been used to predict FGR in recent years. Few studies have been included, and most models are logistic regression, so it is impossible to compare the detection performance of different model types; (2) Due to the lack of the gold standard of FGR, the definition of FGR in different studies is different, so there is a significant bias in the process of model construction; (3) For the incidence of FGR, the number of samples in a single center is small, and the number of samples is not enough to support external verification.
Conclusion

In conclusion, machine learning has an excellent predictive value for FGR, which indicates that practical machine learning can be used as a potential means for the early identification of FGR. Still, it needs to be improved in terms of quality and study design. Therefore, we look forward to harmonizing definitions, developing scoring tools based on cross-racial, multicenter machine learning, and conducting cross-racial, multicenter studies to guide early risk stratification of FGR to identify FGR accurately and intervene in time.

Declarations

Ethics approval and consent to participate

All analyses were based on previous published studies, thus no ethical approval and patient consent are required.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors acknowledge the financial assistance provided by the Fund program: Chinese Multi-Centered Clinical Trial (ChiCTR-SOC-17010976); Suzhou Gusu Health Talents Program (GSWS2019006, GSWS2020055); Jiangsu Provincial Maternal and Child Health Scientific Project (F202044); Suzhou “Rejuvenating Health through Science and Education” Youth Science Project (KJXW2021032); Scientific Program from Gusu School, Nanjing Medical University (GSKY20210232); Case-cohort study from Gusu School, Nanjing Medical University (GSKY20220408)

Author Contribution Statement

Zheng Chenhan and Ji Chunya wrote the main manuscript and fully participated in all analyses. Yin Linliang, Wang Benjing and Deng Xuedong contributed to the study concept and design. Lin Guimei, Zheng Chenhan and Ji Chunya participated in literature search, data extraction, and quality assessment. All authors read and approved the final manuscript.
Acknowledgements

We would like to thank the researchers and study participants for their contributions.

References


**Tables**

Table 1 and 2 are available in the Supplementary Files section.

**Figures**
Identification of studies via databases and registers

Figure 1

Flowchart showing selection of studies for systematic review and meta-analysis
Figure 2

Forest plots of risk of: (a) C-index of training set and validation set; (b) Sensitivity and specificity of the training set; (c) Sensitivity and specificity of the validation set; (d) The c-index defined with DP; (e) The c-index defined without DP; (f) Sensitivity and specificity using DP; (g) Sensitivity and specificity without DP

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.
- Table1.xlsx
- Table2.docx
- TableS1.xlsx
- TableS1.xlsx