Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction: a systematic review

Giulia Zamagni, MSc; Camilla Fregona, MD; Moira Barbieri, MD; Maria Sole Scalia, MD; Lorenzo Monasta, DSc; Christoph Lees, MD, PhD; Tamara Stampalija, MD, PhD; Giulia Barbati, PhD

OBJECTIVES: Fetal growth restriction (FGR) significantly contribute to perinatal morbidity, mortality, and long-term adverse health outcomes. While small for gestational age (SGA) is often used as a proxy for FGR, it does not necessarily indicate pathological growth restriction. Given the increasing interest in machine learning (ML) for predicting FGR/SGA, this study systematically reviews ML applications in this domain, evaluating their methodological rigor and reporting quality, following standardized guidelines.

DATA SOURCES: The systematic search was conducted in MEDLINE and Scopus on June 21, 2024, following PRISMA 2020 guidelines.

STUDY ELIGIBILITY CRITERIA: Eligible studies implemented ML models for FGR/SGA prediction using routinely available clinical variables and reported at least one area under the receiver operating characteristic (AUROC) and/or accuracy. Exclusions included preprints, conference abstracts, systematic reviews, animal studies, and models relying exclusively on biomarkers or genomics, as not part of the clinical practice.

STUDY APPRAISAL AND SYNTHESIS METHODS: Two independent reviewers screened articles with the help of the Rayyan software. Risk of bias was assessed using the PROBAST checklist. Adherence to the guidelines on the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis+artificial intelligence (TRIPOD+AI) was evaluated across methods, results, and discussion sections using a 4-point Likert scale. Sample size adequacy was assessed for each study, accounting for outcome type, predictors, and outcome prevalence.

RESULTS: The search identified 272 studies, with 20 meeting the inclusion criteria. Definitions of FGR/SGA were inconsistent, particularly in technical journals. Adherence to TRIPOD+AI guidelines was variable, as no model reported on fairness or heterogeneity across relevant subgroups, and only 15% reported on calibration. Only 30% of studies met the minimum sample size required for ML models, indicating potential overfitting and limited generalizability.

CONCLUSION: Despite the potential of ML models in predicting FGR/SGA, key limitations persist, including inconsistent outcome definitions, underpowered models, and suboptimal reporting of calibration and clinical applicability. Future studies should emphasize standardized definitions, robust sample sizes, and comprehensive reporting to enhance model reliability and clinical translation.

Key words: artificial intelligence, fetal growth restriction, literature review, small for gestational age

Cite this article as: Zamagni G, Fregona C, Barbieri M, et al. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction: a systematic review. Am J Obstet Gynecol MFM 2026;8:101862.

From the University of Trieste, Trieste, Italy (Zamagni); Clinical Epidemiology and Public Health Research Unit, Institute for Maternal and Child Health – IRCCS "Burlo Garofolo", Trieste, Italy (Zamagni and Monasta); Unit of Fetal Medicine and Prenatal Diagnosis, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy (Fregona, Scalia, and Stampalija); Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy (Stampalija); Unit of Obstetrics and Maternal Fetal Medicine, Department of Women, Newborns and Children, Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, University of Milan, Milan, Italy (Barbieri); Centre for Fetal Care, Queen Charlotte's and Chelsea Hospital, Imperial College Healthcare NHS Trust, London, UK (Lees); Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK (Lees); Biostatistics Unit, Department of Medical Sciences, University of Trieste, Trieste, Italy (Barbati).

Received June 10, 2025; revised November 13, 2025; accepted November 20, 2025.

T.S. and G.B. are joint senior authors.

Condensation: Machine learning models for predicting fetal growth restriction and small-for-gestational-age show inconsistent definitions, underpowered samples, and poor adherence to TRIPOD+AI reporting standards, limiting clinical applicability.

Funding: This research received no external funding.

Data Availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: We declare no competing interests.

Corresponding author: Giulia Zamagni giulia.zamagni@burlo.trieste.it

2589-9333/$36.00

http://dx.doi.org/10.1016/j.ajogmf.2025.101862

AJOG MFM at a Glance

Why was this study conducted?

This systematic review evaluates the methodological rigor and reporting quality of machine learning models developed to predict fetal growth restriction and small-for-gestational-age.

Key findings

Outcome definitions were inconsistent, most models were underpowered, and adherence to reporting standards was suboptimal, particularly regarding calibration, fairness, and generalizability.

What does this add to what is known?

This study highlights the need for standardized definitions, adequate sample sizes, and transparent reporting to enhance the reliability, transparency, and clinical translation of machine learning models in perinatal medicine.

Introduction

Fetal growth restriction (FGR) and small for gestational age (SGA) are one of the major factors contributing to perinatal morbidity and mortality, as well as long-term health consequences, including adverse neurodevelopmental outcomes.[1–3] Although the terms FGR and SGA are often used interchangeably in the medical literature, FGR typically describes the failure of the fetus to achieve its growth potential, whereas SGA refers prenatally to a fetus whose estimated fetal weight or abdominal circumference falls below the 10th percentile of the standard growth charts, and postnatally as a birth weight below the 10th percentile.[4–6] Therefore, as the SGA definition relies solely on biometric measures and does not account for in-utero growth nor Doppler parameters, its application may include also constitutionally small but healthy infants.[7]

Despite a consensus on the diagnosis of FGR was reached in 2016,[8] including both biometric and functional parameters, its clinical utility has recently been questioned, and SGA remains frequently used as a proxy for FGR.[9–11] Predicting FGR/SGA poses significant challenges, yet it is crucial for enabling timely interventions that can enhance perinatal outcomes. Recently, machine learning (ML) techniques have gained popularity in clinical research for their potential ability to manage more complex relationships than the traditional statistical methods, if these approaches are developed and evaluated with rigorous methodologies and within a robust statistical framework.[12]

The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement provides guidelines to improve the methodological quality and reporting of predictive models.[13] With the increasing use of artificial intelligence (AI) technologies, these guidelines have been revised to create the TRIPOD+AI extension, specifically designed to address unique challenges associated with ML models, such as ensuring transparency in model development, evaluation, and generalizability.[14]

Therefore, this study aims to systematically review the scientific literature on ML applications for prediction and definition of SGA/FGR, and to provide the current state of the art concerning methodological rigor.

Methods

Search strategy and eligibility criteria

The systematic search was prospectively registered in PROSPERO (ID CRD42024585914) and conducted following the PRISMA 2020 guidelines,[15] in MEDLINE and Scopus on June 21, 2024. A customized search strategy was employed to ensure a balance between precision and comprehensiveness, considering the distinct characteristics of each database (Supplementary Table 1).

The two queries submitted to MEDLINE and Scopus were equivalent in terms of research objectives, but additional filters were applied in Scopus to narrow the research area and ensure the relevance of the selected studies. On the other hand, in MEDLINE, no filters on the subject area were needed due to its inherent focus on biomedical literature. No limitations were set on the publication date, while only articles in English language were considered eligible. Preprints, conference abstracts, systematic reviews, and animal studies were excluded as they did not align with the objectives of this work. Additionally, inclusion required models reporting at least one performance measure for the final model (eg, area under the receiver operating characteristic [AUROC], accuracy). To ensure clinical applicability, we included only models using routinely available clinical variables (eg, maternal characteristics, fetal biometry, Doppler indices). Therefore, models based exclusively on biomarkers, genomics, or other laboratory-based parameters were excluded, as they may not be widely accessible in routine obstetric practice.

Data extraction

After obtaining the search results, EndNote[16] software was utilized to identify and remove duplicates, as it offers advanced and customizable duplicate detection features. Titles, abstracts, and main text of each selected article were independently screened by two authors (GZ and CF). To streamline the screening process, Rayyan software[17] was utilized to highlight inclusion and exclusion criteria within titles and abstracts. Final decisions were reached through consensus between the two reviewers, ensuring thorough analysis and critical evaluation of each study. This semi-automated approach helped optimize the process while maintaining high precision and consistency.

Risk of bias assessment

The risk of bias assessment for each included study was conducted using the PROBAST checklist, independently completed by two reviewers. Discrepancies were then discussed, and a final consensus judgment is summarized in Supplementary Table 2.

Data synthesis

The adequacy of the sample size adopted in each study was evaluated considering the results of Riley[18] and van der Ploeg.[19]

As a first step, the absolute minimum sample size required in each study was calculated using Riley's formula for clinical prediction models with binary outcomes (ie, logistic regression). It is known that ML methods inherently demand more data compared to traditional statistical techniques.[20] Consequently, Riley's formula establishes a "foundational minimum," and additional data is assumed to be necessary to meet the stringent requirements of the ML approach. Specifically, this minimum sample size was calculated using R package "pmsampsize" and the following command: pmsampsize (type="b," parameters=*N*, prevalence=*P*, shrinkage=0.9, cstatistic=0.8), where "b" indicates a binary outcome, *N* represents the number of candidate predictor parameters, and *P* the outcome prevalence. All the estimated effects were shrunk by 10% to prevent overfitting, and the minimum desirable value for model performance, expressed as c-statistic, was set to 0.8. When outcome prevalence in the target population was not reported, a value of 0.5 was assumed. For example, assuming 20 continuous candidate predictors, an outcome prevalence of 20%, and a target c-statistic of 0.8, Riley's formula yields a minimum of 887 subjects for logistic regression. However, modern ML algorithms (such as random forests, support vector machines [SVM], and neural networks) are more data-intensive and prone to overfitting compared to logistic regression, especially as model complexity increases. As a result, they require substantially larger sample sizes to achieve similar levels of robustness. Therefore, the sample size thresholds presented in our review should be interpreted as conservative lower bounds.

Moreover, adherence to the TRIPOD +AI statement was examined for all the selected papers across the following domains: methods (25 items), results (8 items), and discussion (5 items). The assessment was conducted using a 4-point Likert scale with the following ratings: 1=not adherent (absent/not reported), 2=poorly adherent (the item is present but lacks sufficient methodological details), 3=mostly adherent (the item is present with sufficiently explained methodological details), and 4=adherent (the item is present with all methodological details fully explained). For example, to be rated as adherent in the adequate reporting of performance measures (discrimination, calibration, and clinical utility), at least two out of the three measures must be reported.

For each study, the median with interquartile range and the sum of scores were reported for each domain (ie, "Methods," "Results," "Discussion").

Differences in total scores obtained in each domain were investigated by journal area (ie, technical vs clinical) and evaluated using the Wilcoxon Mann −Whitney test. Correlations between total scores across different domains were investigated using Spearman or Pearson correlation coefficient, as appropriate.

Statistical significance was set at alpha=0.05. All the analyses were conducted using R.[21]

Results

Study selection, study characteristics, and risk of bias

The search strategy identified 272 records published between 1997 and 2024, of which 20 were deemed eligible and included in the analysis. The included studies are listed in Table 1, while the excluded studies are reported in Supplementary Table 3.

The flow diagram illustrating the systematic search is shown in Figure 1.

The distribution of the risk of bias across PROBAST domains is shown in Supplementary Figure 1.

Synthesis of the results

FGR/SGA definition and journal area. In general, two distinct frameworks were identified. The first was aimed at detecting SGA and/or FGR during pregnancy (11 studies), emphasizing early identification and monitoring of these conditions before birth to enable adequate management. The second framework was centered on predicting the SGA/FGR condition at birth (9 studies), guiding postnatal care.

Overall, 10 studies were published in clinical journals and 10 on technical journals. As shown in Table 2, FGR/SGA were not always clearly defined, and even when definitions were provided, they varied significantly among different studies, especially when referring to the prenatal diagnosis. Among these latter studies, most were published after the 2016 Delphi consensus on FGR (90.1%), and four adhered to the Delphi criteria (36.4%). The clarity of FGR/SGA definitions varied also across different journal areas. In particular, the technical field appeared to be less focused on the clinical definitions (Supplementary Figure 2).

Adherence to TRIPOD+AI statement

Adherence to each item of the TRIPOD +AI statement, expressed as a 4-point Likert scale, is shown in Table 3.

Methods

As regards the methods section, the minimum achievable total score was 25, and the maximum was 100. In our set of studies, total scores ranged from 28 to 58 (Supplementary Table 4).

No differences between scores of studies published in clinical journals and those published on technical journals were found (Figure 2, A).

The most problematic items, for which all studies received a score of 1, were the following: 12d ("*Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters [*eg*, hospitals, countries]"*), 14 ("*Describe any approaches that were used to address model fairness and their rationale*"), and 16 ("*Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors*"). Notably, these items were all focused on managing variability and ensuring robustness, fairness, and generalizability across different contexts.

TABLE 1
Studies considered for the systematic review

ID	Author	Title
1	Gurgen (1997)	IUGR detection by ultrasonographic examinations using neural networks
2	Li (2016)	Comparison of different machine learning approaches to predict small for gestational age infants
3	Kuhle (2018)	Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study
4	Signorini (2020)	Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring
5	Crockart (2021)	Classification of intrauterine growth restriction at 34−38 weeks gestation with machine learning models
6	Nguyen-Van (2021)	Identification of latent risk clinical attributes for children born under IUGR condition using machine learning techniques
7	Odendaal (2021).	Accelerations of the fetal heart rate in the screening for fetal growth restriction at 34−38 week's gestation
8	Pini (2021)	A machine learning approach to monitor the emergence of late intrauterine growth restriction
9	Saw (2021)	Machine learning improves early prediction of small-for-gestational-age births and reveals nuchal fold thickness as unexpected predictor
10	Tao (2021)	Fetal birthweight prediction with measured data by a temporal machine learning method
11	Aslam (2022)	Explainable computational intelligence model for antepartum fetal monitoring to predict the risk of IUGR
12	Deval (2022)	A machine learning−based intrauterine growth restriction (IUGR) prediction model for newborns
13	Dieste-Perez (2022)	Personalized model to predict small for gestational age at delivery using fetal biometrics, maternal characteristics, and pregnancy biomarkers: a retrospective cohort study of births assisted at a Spanish hospital
14	Gomez-Jemes (2022)	Machine learning to predict pre-eclampsia and intrauterine growth restriction in pregnant women
15	Li (2023)	A new XGBoost algorithm based prediction model for fetal growth restriction in patients with preeclampsia
16	Song (2023)	Predicting the risk of fetal growth restriction by radiomics analysis of the placenta on T2WI: a retrospective case-control study
17	Taeidi (2023)	Machine learning-based approach to predict intrauterine growth restriction
18	Ruikun Li (2024)	A hybrid model for fetal growth restriction assessment by automatic placental radiomics on T2-weighted MRI and multifeature fusion
19	Vasilache (2024)	Prediction of intrauterine growth restriction and preeclampsia using machine learning-based algorithms: a prospective study
20	Sufriyana (2024)	Prognosticating fetal growth restriction and small for gestational age by medical history

Zamagni. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction. Am J Obstet Gynecol MFM 2025.

Conversely, the highest total scores (ie, 76 and 72, respectively) were found for two items emphasizing the importance of a methodological approach in handling predictors: item 9a ("*Describe the choice of initial predictors [eg, literature, previous models, all available predictors] and any preselection of predictors before model building*") and item 12b ("*Depending on the type of model, describe how predictors were handled in the analyses [functional form, rescaling, transformation, or any standardization]*").

Results

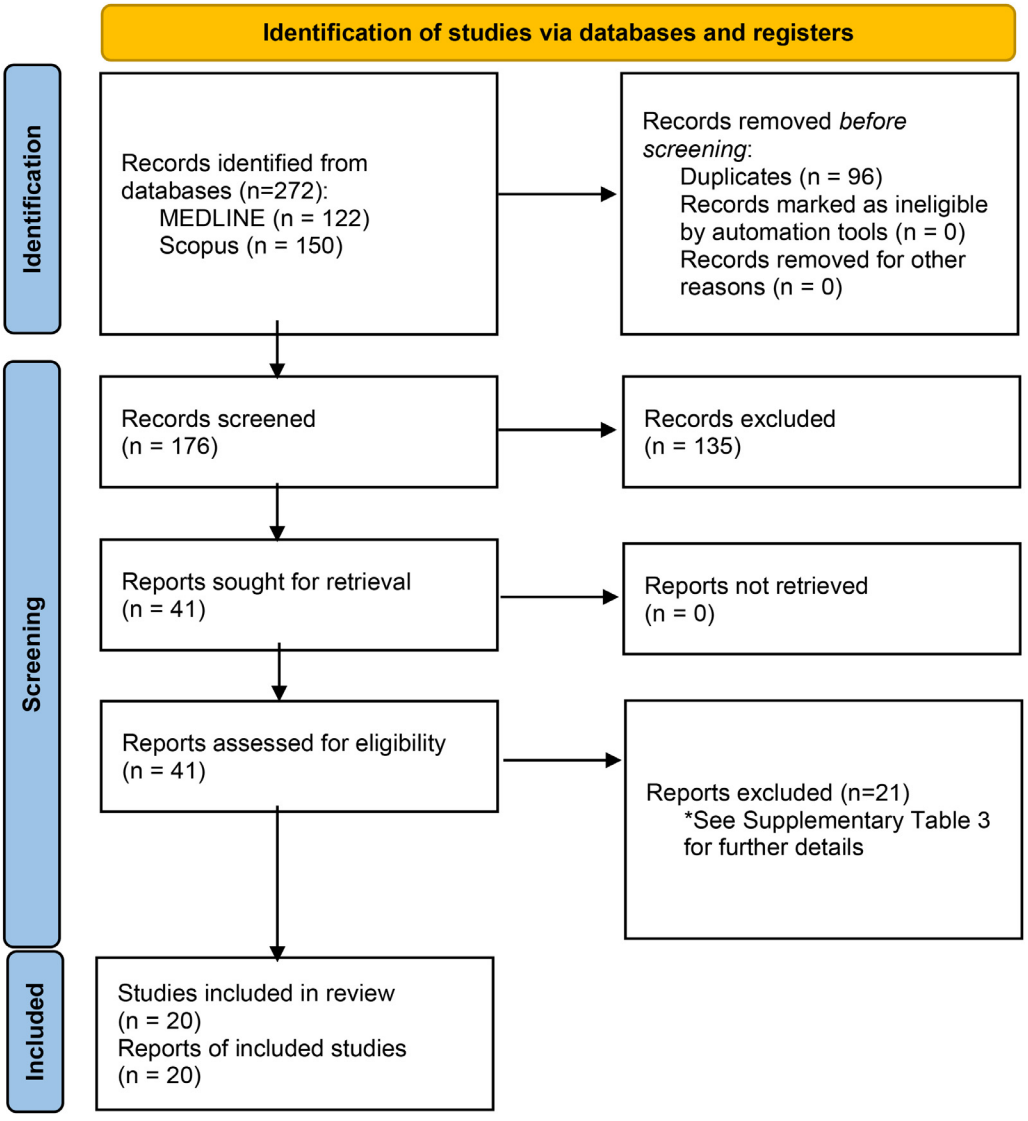
Regarding the results section, the achievable total score ranged from 8 to 32. Despite this range, the selected studies achieved a minimum score of 9 and a maximum of only 18, corresponding to 56.3% of the maximum achievable score (Supplementary Table 4).

As shown in Figure 2, B, studies published in clinical journals showed significantly higher scores than those published in technical journals (P=.029).

Overall, three items sharing a common focus on assessing variability and heterogeneity across different populations or clusters, and ensuring transparent reporting, yielded the most concerning results, each receiving a rating of 1 across all the studies considered: item 20c ("*For model evaluation, show a comparison with the development data of the distribution of important predictors [demographics, predictors, and outcome]*"), item 23b ("*If examined, report results of any heterogeneity in model performance across clusters*") and item 24 ("*Report the results from any model updating, including the updated model and subsequent performance*").

On the other hand, more emphasis was put on item 20a ("*Describe the flow*

FIGURE 1
PRISMA flow diagram illustrating the study selection process



Zamagni. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction. Am J Obstet Gynecol MFM 2025.

of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time").

Item 23a required special attention. In this case, none of the studies received a rating of 1, as reporting performance measures was a prerequisite for inclusion in this review. Nevertheless, most studies did not provide confidence intervals for their performance measures, and none evaluated model performance across different sociodemographic groups, suggesting a lack of robust and equitable model assessment.

In addition, only three studies out of 20 (15%) evaluated model calibration, highlighting how this aspect is often overlooked, with most studies focusing predominantly on AUROC or accuracy.

Discussion

For the discussion section, achievable total scores ranged from 5 to 20. In our set of studies, the range of actual total scores varied between 5 and 14, with five studies (25%) reaching the level of 14 (Supplementary Table 4).

However, studies published in clinical journals received significantly higher scores than those published in technical journals (P=.018, Figure 2, C).

Two items were identified as the most critical, as they reached a score of 1 in all studies, meaning that these aspects were not addressed at all. These items were: "*Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users,*" and "*Describe how poor quality or unavailable input data* (eg, *predictor values)*

TABLE 2
Characteristics of the included studies

ID	Journal area	Sample size	Outcome definition	Outcome prevalence	Number of candidate predictors	Model	Performance metric	Highest performance	Estimated minimum sample size
1	Technical	150	Not clearly defined	NA*	4	NN	Accuracy	0.95	385
2	Technical	215,568	BW<3rd centile	6.2%	342	SVM, RF, LR; Sparse LR	AUROC	0.85	39,416
3	Clinical	30,705	BW<10th centile or BW<3rd centile	7.9%	25	LR, EN, CT, RF, GB, NN	Accuracy; AUROC	0.91; 0.77	2319
4	Technical	120	Abnormal Doppler and BW<10th centile, AC<10th centile and 5-min Apgar score ≤8	50.0%	12	RF	Accuracy	0.91	385
5	Technical	150	EFW<10th centile between 34+0 and 37 +6 wk	13.3%	17	SGD	Accuracy; AUROC	0.77	498
6	Technical	75	Not clearly defined at birth	54.7%	35	LR, XGBoost, SVM	Accuracy	0.95	1030
7	Clinical	4496	EFW<10th centile between 34+0 and 37 +6 wk	7.1%	7	KNN	AUROC	0.81	713
8	Technical	262	Delphi criteria	38.9%	31	SVM	Accuracy	0.85	952
9	Clinical	347	BW<10th centile; BW<3rd centile	58.2%	16	RF, SVM, MLP	Accuracy	0.70; 0.73	479
10	Technical	5759	BW<10th centile	NA*	11	LSTMN	Accuracy	0.93	385
11	Technical	262	Not clearly defined	28.2%	13	RF, SVM, KNN, GB	Accuracy	0.97	463
12	Clinical	214	Not clearly defined at birth	58.9%	31	SMO, MLP, Naive Bayes	Accuracy	0.955	931
13	Clinical	12,912	BW<10th centile	9.9%	19	LR	AURCO	0.86	1458
14	Technical	95	Not clearly defined	12.6%	21	DT, Extra Tree; RF; KNN	Accuracy; AUROC	0.79; 0.87	1318
15	Clinical	303	Delphi criteria	20.5%	27	XGBoost	AUROC	0.85	1177
16	Clinical	202	Delphi criteria	51.5%	23	LR	AUROC	0.87	673
17	Clinical	8683	AC<2×SD+mean	8.2%	18	CT, RF, DL, GB	AUROC	0.91	1618
18	Clinical	274	EFW<10th centile; UA-PI>95th centile; BW<10th centile	45.6%	7	RF	AUROC	0.88	382
19	Clinical	210	Delphi criteria	7.1%	12	DT, Naive Bayes, SVM, RF	Accuracy	0.96	1222
20	Technical	26,576	P05 prefix of ICD-10	0.4%	54	DI-VNN	AUROC	0.74	88,094

AC, abdominal circumference; *ANN*, artificial neural networks; *AUROC*, area under the ROC curve; *BNN*, Bayesian neural networks; *BW*, birthweight; *CPR*, cerebral-placental ratio; *CT*, classification trees; *DI-VNN*, dynamic interaction variable neural network; *DL*, deep learning; *DT*, decision trees; *EDF*, end diastolic flow; *EFW*, estimated fetal weight; *EN*, elastic net; *fCNN*, fully convolutional neural networks; *GB*, gradient boosting; *GLM*, generalized linear model; *KNN*, K-nearest neighbor; *LR*, logistic regression; *LSTMN*, long short-term memory network; *MLP*, multilayer perceptron; *NN*, neural networks; *RF*, random forest; *SGD*, stochastic gradient descent; *SMO*, sequential minimal optimization; *SNF*, similarity network fusion; *SVM*, support vector machine; *UA*, umbilical artery; *UA-PI*, umbilical artery pulsatility index; *UtA-PI*, uterine artery pulsatility index; *XGB*, extreme gradient boosting.

* not available.

Zamagni. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction. Am J Obstet Gynecol MFM 2025.

should be assessed and handled when implementing the prediction model." In contrast, more attention was dedicated to the overall interpretation of the main results, considering previous studies.

Correlation analysis of scale scores

As illustrated in Supplementary Figure 3, a clear pattern emerged, suggesting a moderately consistent level of adherence to TRIPOD+AI guidelines across different parts of the paper.

There was a mild yet statistically significant correlation between the total scores obtained in the methods section and those in the results section (P=.012). A similar positive correlation was observed between the total scores for the "Methods" and "Discussion" sections (P=.031). However, the strongest relationship was noted between the total scores for "Results" and "Discussion" (P=.011).

Modelling techniques and sample size requirements

Modelling strategies and sample sizes employed in each study are shown in Table 1. Overall, the minimum sample size estimated via Riley's formula was achieved by six studies (30%).

For each study that fell short of the minimum sample size, the discrepancy between the actual and the minimum required sample size was computed and expressed as a percentage (Supplementary Figure 4). The median percentage difference was 70% (IQR 61%−77%), indicating a systematic under-power across the analyzed studies.

The model strategy more frequently adopted was SVM, employed in six studies out of 20 (30%).

Comment

Main findings

Our systematic review provides an overview over the methodological rigor and adherence to standardized definitions and methodological guidelines in

TABLE 3

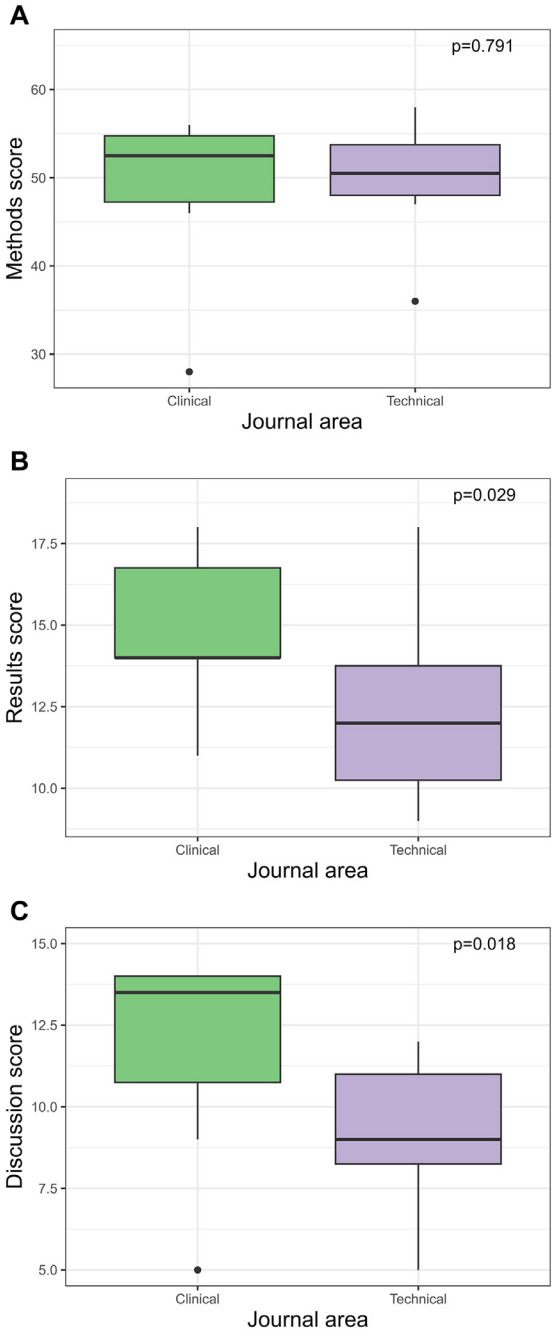
Results of the evaluation of the adherence of each selected study to TRIPOD+AI guidelines (methods, results, and discussion sections)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Item	**Methods**																			
5a	1	3	3	2	3	1	2	2	3	1	3	1	2	3	3	3	3	3	3	3
5b	1	4	4	1	2	1	2	1	1	4	2	1	4	4	4	4	4	4	4	1
6a	1	4	2	4	1	1	1	4	4	1	1	1	4	4	4	4	4	4	4	4
6b	1	4	4	4	1	1	1	4	4	1	4	1	4	1	4	4	4	4	4	4
6c	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	1	3	3	3	3	3	3	3	3	3	3	1	1	3	3	3	1	3	3	1
8a	2	3	3	3	3	3	3	3	3	3	3	2	3	2	3	3	3	3	3	3
8b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8c	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9a	3	4	4	4	4	4	4	4	4	4	4	1	4	4	4	4	4	4	4	4
9b	2	3	3	3	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3
9c	1	1	1	1	3	1	3	1	3	1	1	1	1	1	1	1	1	3	1	1
10	1	1	1	1	2	2	2	2	2	2	2	1	2	1	2	1	1	2	1	2
11	1	4	4	4	1	4	1	4	1	1	1	1	1	4	1	1	1	1	1	1
12a	2	3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3
12b	2	4	4	4	4	4	4	4	4	4	4	1	4	4	4	4	1	4	4	4
12c	3	4	4	4	3	3	3	3	4	3	3	2	2	3	3	2	2	3	2	3
12d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12e	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2	3	2	2	2	4
12f	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12g	4	1	1	2	1	3	1	1	1	3	4	1	1	4	1	1	1	1	1	4
13	1	3	3	1	2	1	1	2	3	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	3	3	3	4	3	3	3	4	3	4	3	3	3	3	3	3	3	3	3	3
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Results																				
20a	1	4	4	4	1	1	4	4	4	1	1	1	4	3	4	4	4	4	4	4
20b	1	1	4	2	1	1	2	2	2	2	1	1	2	1	3	3	1	3	3	3
20c	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1	4	1	1	3	1	1	1	4	1	4	1	1
22	2	1	4	1	1	3	1	1	1	1	1	1	1	4	1	1	1	1	1	4
23a	3	2	2	3	2	3	3	2	2	2	2	2	3	2	2	2	3	2	2	3
23b	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Discussion																				
25	4	4	4	4	4	2	4	4	4	1	2	1	3	4	4	4	4	4	4	1
26	1	1	4	1	1	1	1	1	4	1	1	1	4	3	1	4	4	4	4	1
27a	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27b	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27c	2	4	3	4	2	3	2	4	4	1	4	1	3	4	3	4	4	4	4	2

The dash indicates nonapplicability. Study IDs correspond to those reported in Table 1, where full author names and publication details are provided; Rows correspond to the TRIPOD+AI checklist items, labeled according to the original guideline numbering; Scoring is based on a 1 to 4 scale, where 1=not adherent, 2=poorly adherent, 3=mostly adherent, and 4=fully adherent.

Zamagni. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction. Am J Obstet Gynecol MFM 2025.

FIGURE 2
Boxplots showing the distribution of methods (A), results (B), and discussion (C) scores for clinical and technical journals



Zamagni. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction. Am J Obstet Gynecol MFM 2025.

studies employing ML models for prediction of FGR or SGA.

Adherence to the uniform definition of FGR resulted poor, with most studies relying on SGA as a proxy for FGR, undermining the overall comparability and aggregation of the results.[22,23] Although there is still a debate over the definition of FGR, this represents a main limitation for data comparison and research or clinical consistency.[24] From a clinical perspective, using SGA as a proxy for FGR can lead to inappropriate clinical management and increased maternal anxiety. In particular, healthy, constitutionally small fetuses may be exposed to unnecessary interventions such as increased monitoring, hospital admission, or preterm delivery. Conversely, some fetuses with true growth restriction—who do not meet SGA criteria but show evidence of placental insufficiency (eg, abnormal Doppler findings or growth deceleration)—may go unrecognized, missing opportunities for adequate surveillance and timely intervention. Therefore, clearly distinguishing between SGA and FGR is essential to ensure optimal care and perinatal outcomes.

Our systematic review showed also that studies published in technical journals were less focused on providing clear definitions of FGR/SGA compared to those in clinical journals. This difference might suggest that, in the technical field, methodological details may be prioritized over clinical relevance, potentially impacting the overall applicability of the findings. However, model development should be viewed as a responsible effort focused on clinical impact, rather than solely a technical exercise. Involving clinical experts at every stage, particularly in defining outcomes and predictors, is essential to ensure that prediction models are grounded in clinically meaningful concepts and relevant endpoints.

Regarding the methodological aspects, our analysis revealed that many studies did not adequately address key aspects, especially involving heterogeneity, model fairness, or differences between development and evaluation data. In fact, all studies received low scores on items related to managing variability and ensuring generalizability, revealing a suboptimal focus on the overall clinical applicability. It should be highlighted that only through the collaborative efforts of multidisciplinary teams can a translational clinical risk prediction model be developed, thereby driving advancements in clinical practice.[25] Furthermore, external validation

is crucial to confirm that ML models deliver accurate predictions in different clinical settings and populations. Without this step, models may misclassify risks when used outside their development context, resulting in inappropriate and potentially harmful clinical decisions. In particular, in fetal surveillance and pregnancy management, this may lead to unnecessary interventions for those at low risk or a failure to identify high-risk cases, ultimately impacting counseling as well as maternal and neonatal outcomes. For this reason, any model that has not undergone external validation cannot be considered suitable for clinical use.

The assessment of the results' section highlighted also several gaps, with most studies that failed to report confidence intervals for performance measures or evaluate model performance across different meaningful sociodemographic groups. Reporting uncertainty is essential, as confidence intervals provide critical information on the precision and robustness of model estimates, particularly in clinical settings where predictions inform individual-level decisions. In addition, evaluating model performance across relevant sociodemographic subgroups is crucial to ensure that predictive models perform equitably and do not inadvertently reinforce existing health disparities. Therefore, the lack of a detailed and equitable approach significantly influences the reliability of the results and confirms the need of a clear benchmark of best practices.

In addition, most of the considered studies did not assess model calibration or clinical utility, focusing solely on model classification performance. However, it is well known that a model may demonstrate strong discriminative performance but poor calibration, which can significantly affect its effectiveness, especially in the clinical context, where the risk communication must be clear. Specifically, poor calibration means that the predicted risks do not accurately correspond to actual outcomes, which can lead to both over- and under-treatment. For instance, overestimated risks may prompt unnecessary interventions, increase patient anxiety, or result in the use of additional healthcare resources. Conversely, underestimated risks could delay necessary care, potentially compromising perinatal outcomes.

Furthermore, even a model with good discriminative performance and calibration may not provide a substantial net benefit. Therefore, the implementation of a prediction model in the clinical practice should be carefully evaluated to avoid unnecessary resource deployment.[26]

Regarding the overall reliability, our findings indicated a systematic underpower in the studies considered. ML models, especially those designed for complex tasks such as predicting health outcomes, often require substantial amounts of data to achieve reliable performance and generalize well to new, unseen data. Several issues can arise from underpowered models. First, a study with a small sample size can easily lead to overfitting training data, reducing generalizability to different populations or settings and, therefore, practical utility.[27,28] Moreover, in practical ML applications, releasing an underpowered model can lead to an inefficient use of the available resources and to a suboptimal decision-making process.[29]

Strengths and limitations

This study has several strengths and limitations. A strength is the comprehensive search strategy, which adhered to PRISMA 2020 guidelines and employed customized approach aimed to ensure precision and relevance. This methodology enabled the inclusion of several studies across two databases and different subject areas, enhancing the overall robustness of the findings. Moreover, the use of semi-automated tools for managing and reviewing the studies contributed to a high level of consistency and accuracy in the selection process.

On the other hand, there are some limitations to acknowledge. The retrospective application of the TRIPOD+AI guidelines to studies published before their introduction may not fully reflect the methodological practices at the time of those studies. We acknowledge that this retrospective approach, while useful for identifying temporal trends, may also impose anachronistic standards on earlier research. Furthermore, although our review indicates a modest improvement in methodological rigor and reporting in more recent studies, the number of publications following the release of TRIPOD +AI remains limited, with most included studies published before or within 1 month of the guidelines. As more studies become available in the coming years, future reviews will be better equipped to quantitatively evaluate the impact of these updated reporting standards.

However, despite these limitations, the picture provided by this review can still guide clinicians in understanding the current state of the art in predictive modelling for SGA/FGR, offering insights into both the progress made and the areas requiring further refinement. To overcome current limitations in the application of ML for predicting SGA/FGR, future research should focus on developing larger and more diverse datasets, improving model interpretability, and ensuring external validation across different populations. This goal can be facilitated by establishing multicenter, prospectively collected cohorts that ensure adequate representation of relevant subgroups, thereby enhancing the generalizability and robustness of ML models. Standardizing FGR and SGA definitions, ideally in partnership with professional societies, is also essential for better data clarity and comparability, as well as clinical translation.

Consistent with the TRIPOD+AI guidelines, the systematic assessment of model fairness and calibration should be considered a mandatory component of ML model reporting and evaluation, as both equity and reliability are fundamental for safe clinical application. On the data sharing front, engagement with perinatal data linkage networks, such as the Euro-Peristat initiative,[30] can enable the pooling of heterogeneous cohorts and support external validation efforts. Additionally, the creation of perinatal and obstetric registries, such as the Finnish Medical Birth Register,[31] not only facilitate high-quality data sharing for research, but also could enable surveillance of maternal and children health outcomes. These registries could also provide a foundation for external model validation, longitudinal monitoring of interventions, and ongoing

assessment of both short- and long-term trends in perinatal health.

Finally, collaborative efforts between clinicians and data scientists are essential to design tools that are not only accurate but also clinically meaningful.

Conclusions and implications

In summary, despite the potential of ML models in predicting SGA/FGR, significant challenges persist, especially regarding the reporting of the results, sample size limitations, and discussion quality. Ensuring consistent application of standardized definitions, improving methodological rigor and transparency, and addressing issues of under-powering are critical steps for advancing this research field and enhancing the comparability and applicability of predictive models. ■

CRediT authorship contribution statement

Giulia Zamagni: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Camilla Fregona:** Resources, Formal analysis. **Moira Barbieri:** Writing – review & editing. **Maria Sole Scalia:** Writing – review & editing. **Lorenzo Monasta:** Writing – review & editing, Methodology. **Christoph Lees:** Writing – review & editing. **Tamara Stampalija:** Writing – review & editing, Supervision, Conceptualization. **Giulia Barbati:** Writing – review & editing, Supervision, Methodology, Conceptualization.

ACKNOWLEDGMENTS

This work was supported by the Italian Ministry of Health, through the contribution given to the Institute for Maternal and Child Health IRCCS Burlo Garofolo, Trieste—Italy (Ricerca Corrente RC 08-24).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ajogmf.2025.101862.

REFERENCES

1. Gardosi J, Madurasinghe V, Williams M, et al. Maternal and fetal risk factors for stillbirth: population-based study. BMJ 2013;346(3):f108. https://doi.org/10.1136/bmj.f108.

2. Mendez-Figueroa H, Truong VT, Pedroza C, et al. Small-for-gestational-age infants among uncomplicated pregnancies at term: a secondary analysis of 9 maternal-fetal medicine units network studies. Am J Obstet Gynecol 2016;215(5):628.e1–7. https://doi.org/10.1016/j.ajog.2016.06.043.

3. Arcangeli T, Thilaganathan B, Hooper R, Khan KS, Bhide A. Neurodevelopmental delay in small babies at term: a systematic review. Ultrasound Obstet Gynecol 2012;40(3):267–75. https://doi.org/10.1002/uog.11112.

4. Morris KR, ohnstone JE, Lees C, Morton V, Smith G, Royal College of Obstetricians and Gynaecologists. Investigation and Care of a Small-for-Gestational-Age Fetus and a Growth Restricted Fetus (Green-top Guideline No. 31). BJOG 2024;131(9):e31–80. https://doi.org/10.1111/1471-0528.17814.

5. Lees CC, Romero R, Stampalija T, et al. Clinical opinion: the diagnosis and management of suspected fetal growth restriction: an evidence-based approach. Am J Obstet Gynecol 2022;226(3):366–78. https://doi.org/10.1016/j.ajog.2021.11.1357. Epub 2022 Jan 10. PMID: 35026129; PMCID: PMC9125563.

6. Lees CC, Stampalija T, Baschat A, et al. ISUOG practice guidelines: diagnosis and management of small-for-gestational-age fetus and fetal growth restriction. Ultrasound Obstet Gynecol 2020;56(2):298–312. https://doi.org/10.1002/uog.22134. PMID: 32738107.

7. McCowan LM, Harding JE, Stewart AW. Customized birthweight centiles predict SGA pregnancies with perinatal morbidity. BJOG 2005;112(8):1026–33. https://doi.org/10.1111/j.1471-0528.2005.00656.x.

8. Gordijn SJ, Beune IM, Thilaganathan B, et al. Consensus definition of fetal growth restriction: a Delphi procedure. Ultrasound Obstet Gynecol 2016;48(3):333–9. https://doi.org/10.1002/uog.15884.

9. Society for Maternal-Fetal Medicine (SMFM). Society for Maternal-Fetal Medicine Consult Series #52: diagnosis and management of fetal growth restriction: (Replaces Clinical Guideline Number 3, April 2012). Am J Obstet Gynecol 2020;223(4):B2–17. https://doi.org/10.1016/j.ajog.2020.05.010. Epub 2020 May 12. PMID: 32407785.

10. Alda MG, Holberton J, MacDonald TM, Charlton JK. Small for gestational age at preterm birth identifies adverse neonatal outcomes more reliably than antenatal suspicion of fetal growth restriction. J Matern Fetal Neonatal Med 2023;36(2):2279017. https://doi.org/10.1080/14767058.2023.2279017.

11. Monier I, Ego A, Hocquette A, et al. Validity of a Delphi consensus definition of growth restriction in the newborn for identifying neonatal morbidity. Am J Obstet Gynecol 2025;232(2). https://doi.org/10.1016/j.ajog.2024.04.033. 224.e1–13Epub 2024 Apr 30. PMID: 38697341.

12. Patel MR, Balu S, Pencina MJ. Translating AI for the Clinician. JAMA 2024;332(20):1701–2. https://doi.org/10.1001/jama.2024.21772.

13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13(1):1. https://doi.org/10.1186/s12916-014-0241-z.

14. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.

16. The EndNote Team. EndNote 20 version. Philadelphia, PA: The EndNote Team; Clarivate; 2013 EndNote.

17. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4.

18. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. https://doi.org/10.1136/bmj.m441.

19. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14(1):137. https://doi.org/10.1186/1471-2288-14-137.

20. Infante G, Miceli R, Ambrogi F. Sample size and predictive performance of machine learning methods with survival data: a simulation study. Stat Med 2023;42(30):5657–75. https://doi.org/10.1002/sim.9931. Epub 2023 Nov 10. PMID: 37947168.

21. R Core Team. R: a language and environment for statistical computing [computer program].Vienna, Austria: R Foundation for Statistical Computing; 2024. Available at: https://www.R-project.org/.

22. Fantasia I, Zamagni G, Lees C, et al. Current practice in the diagnosis and management of fetal growth restriction: an international survey. Acta Obstet Gynecol Scand 2022;101(12):1431–9. https://doi.org/10.1111/aogs.14466. Epub 2022 Oct 10. PMID: 36214456; PMCID: PMC9812103.

23. Mylrea-Foley B, Napolitano R, Gordijn S, Wolf H, Lees CC, Stampalija T. TRUFFLE-2 Feasibility Study Authors. Do differences in diagnostic criteria for late fetal growth restriction matter? Am J Obstet Gynecol MFM 2023;5(11):101117. https://doi.org/10.1016/j.ajogmf.2023.101117. Epub 2023 Aug 5. PMID: 37544409.

24. Lees C, Stampalija T, Hecher K. Diagnosis and management of fetal growth restriction: the ISUOG guideline and comparison with the SMFM guideline. Ultrasound Obstet Gynecol 2021;57(6):884–7. https://doi.org/10.1002/uog.23664. PMID: 34077604.

25. Disis ML, Slattery JT. The road we must take: multidisciplinary team science. Sci Transl Med 2010;2:22cm9. https://doi.org/10.1126/scitranslmed.3000421.

26. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6. https://doi.org/10.1136/bmj.i6.

27. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365. https://doi.org/10.1371/journal.pone.0224365.

28. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 1991;13(3):252–64. https://doi.org/10.1109/34.75512.

29. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2.

30. Euro-Peristat: monitoring and evaluating perinatal health in Europe. Available at: https://www.europeristat.com. Accessed on: November 2025.

31. Finnish Institute for Health and Welfare. Medical Birth Register. Available at: https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions/newborns. Accessed on: November 2025.