RESEARCH PAPER

Human-Centred Trust in AI-Supported Risk Communication for High-Risk Pregnancy Care

Wang Chunxiang[1], Ali Napiah Nasution[2], Liena[3]
[1-3]Department of Obstetrics Universitas Prima Indonesia
Co Author Email: alinapiahnasution@unprimdn.ac.id

ABSTRACT
Digital prediction tools are progressively being utilized in the field of high-risk pregnancy care to diagnose preeclampsia, gestational diabetes and preterm birth and intrapartum fetal compromise. However, discrimination statistics are not the sole determinants of the clinical value of artificial intelligence. Development of trust in risk communication is based on how a prediction is explained, verified, governed and connected to collective action. This is a secondary fast review that analyzes the human-centred trust in AI-supported risk communication during high-risk pregnancy care. Thematic analysis coding was the primary means of coding evidence across patient-facing pregnancy digital tools, preeclampsia prediction, prediction of gestational diabetes, preterm prediction and cardiotocography/acidemia detection. Four themes were created: a high level of trust exists when AI assists rather than substitutes for clinicians; explainability and communication translate risk scores into actionable interpretations; prediction models have excellent potential but weak methodology; and equity, privacy and workflow suggest trust to levels beyond AUROC. This review concludes that AI can reinforce risk communication through providing earlier review, personalised counselling and regular escalation. Prospective impact, calibration, subgroup performance, clinician override, anxiety effects and patient understanding are not well reported in model-centred evidence, however. It is not an anti-AI argument based on the argument here. Instead, responsible obstetric AI needs to be developed as a communication and care-relationship technology. Human-centered implementation, as part of it, should include local validation, explainable output, clinician accountability, privacy protection, an equity audit, and plain-language shared decision-making.

Keywords: artificial intelligence; high-risk pregnancy; risk communication; trust; obstetrics; preeclampsia; gestational diabetes; preterm birth; thematic analysis
How to cite this article: Chunxiang W, Nasution AN, Liena. Human-Centred Trust in AI-Supported Risk Communication for High-Risk Pregnancy Care. Int J Drug Deliv Technol. 2026;16(56s): 842-853. DOI: 10.25258/ijddt.16.56s.91
Source of support: Nil.
Conflict of interest: None.

INTRODUCTION

Before it is a technology problem, it is high-risk pregnancy care, which is a trust-sensitive mode of risk communication. A pregnant woman who is informed that she might get preeclampsia, gestational diabetes, preterm birth or fetal compromise is not being given a probability. Uncertainty is demanding the reinterpretation of her body, her fetus, her future appointments and in some instances her birth plan. The area is now being penetrated by artificial intelligence (AI) due to the creation of large amounts of data in routine and intrapartum antenatal and signal-based care. Machine learning systems can analyse electronic health records, lab results, ultrasound images, blood pressure readings, cardiotocography images, and even data from health monitoring apps. The guarantee is the prior acknowledgement of the possibility of preventable damage. The danger is that women will get opaque, poorly framed or provocative outputs that undermine, not upgrade, shared care.

There are high stakes. Statistics on maternal health worldwide, a great deal of which has been compiled, reveal that much remains to be done. It is estimated that about 260,000 women died due to maternal causes in 2023, and the World Health Organization estimated that there were over 700 deaths each day in 2023 of women caused by preventable childbirth and pregnancy-related reasons (UNICEF, 2025; World Health Organization [WHO], 2025a). It is not exceptional cases but high-risk conditions. Globally, preeclampsia occurs in 3-8% of pregnancies and is one of the top causes of maternal and neonatal morbidity (American College of Obstetricians and Gynecologists [ACOG], 2020; WHO, 2025b). The prevalence of gestational diabetes is about 1 in 7 pregnancies worldwide, and estimates vary depending on the population and diagnostic criteria (International Diabetes Federation [IDF], 2025; Wang et al., 2021). Preterm birth is also significant: WHO estimates that in 2020, 13.4 million preterm babies were born, or nearly a tenth of all live births (WHO, 2023). These statistics are a plausible entry point for community health into AI-powered risk communication. When predictive tools can recognize the threat at an earlier stage and assist in

communicating out-of-action plans to clinicians, they can enhance prevention, surveillance and timely referral.

However, risk prediction is not the same as risk communication. One model can have a high AUROC value. Still, patients might not interpret the value, the reasons the score changed, what data was used, whether the model is effective for someone like them, or who is responsible in the case of a prediction error. Moral and emotional complexity increases these concerns in pregnancy. Motions can influence the well-being of the mother and fetus, including relatives and result in interventions like aspirin prophylaxis, glucose checks, hospitalization, induction, caesarean section or neonatal intensive care planning. A false negative can delay care. False positive can add burden, stress and intervention to surveillance. Reliance by humans is thus an appropriate form of trust and not blind trust.

This article embraces the terminology of human-centre trust to imply a justified belief that AI-intensified risk communication will be clinically safe, intelligible, supervised, respectful and fair. This is unlike mere technology acceptance. A woman might believe an app because of its convenience, but still mistrusts a risk score provided by an algorithm without interpretation or validation by a clinician. On the other hand, she may be afraid of AI. Still, she can accept it when her midwife or obstetrician assures her that it is a second opinion, not a substitute for professional judgement. The review question is: How can AI-assisted risk communication in high-risk pregnancy practice foster human-centred trust and safeguard safety, equity and patient agency?

This research cross-tabulates multiple high-risk pregnancy areas rather than treating obstetric AI as a homogeneous unit. Third, it is critical of both methodology and outputs, in that clinical trust is based on design decisions, including retrospective versus prospective validation, external validation, interpretability, patient participation, workflow integration and subgroup testing. The algorithm alone does not generate trust as depicted in Figure 1. It arises when clinical information, AI estimates, clinician evaluations, plain-language communication, and governance defenses interact collectively.

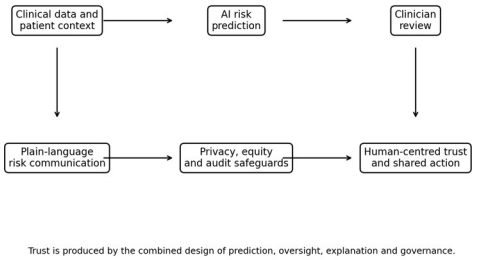


Figure 1. Human-centred trust framework for AI-supported risk communication.

Literature Review and Theoretical Positioning

The literature review is based on secondary reviews, conceptual theory and clinical guidance and not on the primary studies, which will be coded later within the findings. This distinction is significant since the critique of AI in maternal and fetal health frequently summarises a blend of model development papers, feasibility reports and commentaries, which may confuse the evidence and its interpretation. Reviews indicate that there is increasing interest in individualised maternal and fetal AI, including screening, monitoring and decision support, but caution that most models are still in the early stages, are technically heterogeneous and are not yet in regular practice (Correia et al., 2025; Davidson et al., 2025). Equally, machine learning has been written about in systematic reviews of preeclampsia and cardiotocography, with promising measures of prediction; however, reporting is inconsistent, definitions of outcomes vary, and ultimately, outcomes are not prospectively validated (Ahmed et al., 2024; Petrozziello et al., 2019).

The initial lens is trust concerning automation theory. According to Lee and See (2004), trust ought to be tuned towards the purpose, process and performance of a system. Mistrust can prevent the adoption of useful support under-trust, whereas it can cause automobile bias and a lack of vigilance under-trust. The two errors are possible in obstetrics. One clinician might choose to ignore an honest alert due to the tool being opaque or intrusive, and another might over-trust an ostensibly accurate score without verifying whether the patient is different from the development population. Confidence in automation thus endorses a monitored pattern of AI, whereby clinicians retain the decision, explain results and clarify uncertainty to patients.

One of the second lenses follows the technology acceptance theory. Towards this end, Davis (1989) and subsequent UTAUT studies state that perceived usefulness, ease of use, social influence and facilitating conditions influence adoption (Venkatesh et al., 2012). These are applicable yet incomplete

concepts to high-risk pregnancy care. A risk score might benefit the hospital and be something clinicians can conveniently find, but a patient might feel threatened or alienated. That is why the use of patient-centred care and shared decision-making is required as some supplementary frameworks. Values should be respected, information provided clearly, and emotional support and involvement in decision-making are imperative in patient-centred care (Institute of Medicine, 2001). Communication should also not make the woman a risk category in high risk pregnancy. It ought to make her aware of the risk, the options available, the trade-offs involved, and the care she will react to.

The theory of risk perception enriches this. Slovic (1987) demonstrated that people's responses to hazards were driven by dread, control, familiarity, catastrophic potential and trust in institutions rather than by statistical probability. Numeric accuracy is not the only influence on AI-assisted pregnancy risk communication, then. A woman can react in different ways to the 20% of preeclampsia risk based on whether there is a past trauma, whether they know what preeclampsia is, whether they feel in control of the hospital or whether they feel that the number of the population computed the score. Good risk communication should articulate uncertainty without amplifying fear. Timely surveillance, joint planning and guidance have already been highlighted in clinical guidelines on hypertension and diabetes in pregnancy (ACOG, 2020; NICE, 2020, 2023). Instead of opening up a parallel black-box channel, AI ought to enhance these procedures.

The solution to distrust is often proposed as explainable AI; however, the literature is skeptical. Miller (2019) asserts that useful explanation is social: we desire contrastive, selective and actionable explanations, and not the weights of technical features. Health-care reviews also indicate that explainability can positively affect clinician trust but also generate false reassurance when visually compelling but clinically inaccurate explanations are provided (Rosenbacke et al., 2024). In the case of pregnancy care, SHAP plots, risk factor rankings, or saliency maps are insufficient. They should be converted into communication that is relevant to patients, e.g., what modifiable factors exist, how monitoring will adapt, which red flags to be attentive to, and when to seek emergency care.

Moreover, lastly, responsible AI literature focuses on privacy, accountability, fairness and governance (Char et al., 2018; WHO, 2021). These are the main principles of pregnancy since reproductive information is very sensitive and inequitable models may serve to increase existing disproportions. Maternal risk is defined by ethnicity, socioeconomic status, geography, comorbidity, access to care, language, disability and previous obstetric history. When AI models are trained on small groups or poorly recorded data, they can give a false sense of security to underrepresented groups. This is why the literature upholds a human-centred model whereby trust is based on performance; however, it is also on representation, supervision, explanation, acceptance and constant audit.

METHODS

The secondary rapid review design was employed. It was secondary since no recruitment or additional clinical data were collected. The review used an adapted PRISMA logic for transparent searching and screening. However, it was not formatted as a complete systematic review because the aim was an interpretive thematic synthesis rather than a comprehensive meta-analysis (Page et al., 2021). The approach involved integrating structured searching, eligibility criteria, primary-study extraction, critical appraisal and thematic analysis.

The sources of literature reviews were not restricted and included secondary reviews, theories, ethical bases, clinical guidelines and conceptual papers. The following were sources that were used to frame trust, risk communication, AI governance and high-risk pregnancy care. They, however, were not coded as findings. It was chosen to limit the findings section to primary empirical studies, such as qualitative studies, mixed-methods studies, retrospective and prospective cohort model-development studies, external validation studies and reader or signal-analysis studies in pregnancy-related AI or digital monitoring. This was done to minimize circularity. It does not allow a review to take the interpretations of other authors as though they are new empirical results.

They theoretically grouped the searches based on five empirical domains: patient/pregnant-person trust in AI and digital monitoring; predicting preeclampsia, gestational diabetes, and preterm birth; and predicting intrapartum fetal monitoring or acidemia. The eligible studies were primary research, included pregnancy, fetal monitoring or antenatal risk and reported data pertinent to prediction, digital risk communication, patient trust, clinical interpretability or implementation. Findings were not considered in cases where the studies were editorials, protocols, reviews, theoretical papers, or were merely technical reports of no interest in pregnancy and secondary summaries. The last thematic results were based on 20 primary studies, published predominantly between 2020 and 2026. The emphasis on recent research was appropriate, as the AI approach to obstetrics and the demands of responsible AI have evolved rapidly.

The data extraction included study domain field, design, sample or datasets, AI or digital process,

outcome, relevance of trust or communication and methodology limitations. The methodological critique assessed whether the studies were prospective or retrospective, whether they had external validation, whether they reported calibration and subgroup performance, whether they reported patient or clinician communication, and whether the models were related to the workflow. The importance of this is that retrospective data cannot show that a model that works well in high-risk pregnancy care can be explained and operationalised safely in an antenatal clinic.

Thematic analysis was adapted from the reflexive approach of Braun and Clarke, which was designed to combine documents and synthesize evidence (Braun and Clarke, 2006, 2021). The first codes were clinician oversight, patient explanation, privacy, model validation, calibration, external generalisability, subgroup bias, workflow burden, false positives, false negatives, escalation, and anxiety. The comparison of codes was conducted across domains, followed by combining the codes into four analytical themes. The counting of the themes was not mechanical. Counts were, rather, supportive in their nature as a descriptive tool, and the interpretation was made on how evidence, as gathered through various research methods, either concurred or did not. The distribution of primary studies per domain is presented in Figure 2, and the relative intensity of thematic coding across domains is presented in Figure 3.

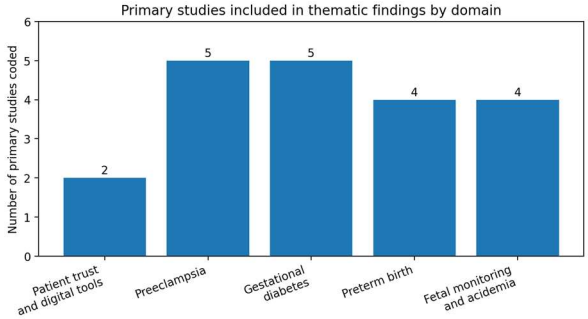


Figure 2. Distribution of primary studies coded in the thematic analysis.

As Figure 2 demonstrates, primary evidence was concentrated around both preeclampsia and gestational diabetes, with fewer studies directly aimed at patient trust or communication via the app. This asymmetry is important since a human-centred trust review can not be based solely on model measurements. An acceptable discriminatory preeclampsia model still can fail, when it develops an unexplainable high-risk imprint which a woman perceives as an accusation or as a fright. Likewise, a fetal monitoring algorithm can enhance the recognition of acidemia. However, it may also increase the number of caesarean procedures, as clinicians and patients may perceive alerts as orders rather than as decision-support information.

FINDINGS

AI as Clinical Support, Not Replacement

AI earns the most trust when it assists clinicians rather than replaces them. This was consistent with the patient-facing studies. Research by Oprescu et al. (2022) demonstrated that pregnant women were interested in AI when it was responsible, useful, safe and explainable. However, the lack of trust and privacy issues decreased enthusiasm. Similar cases involving communication and monitoring tools were observed by Lazarevic et al. (2023). However, this interest was also linked to continuity, reassurance, and access to professionals, rather than independent self-diagnosis. Such discoveries are consistent with and refine the technical studies. Li et al. (2024) and Eberhard et al. (2025) applied interpretable or longitudinal modelling to enable repeated clinical prediction in preeclampsia, but neither study showed patient-facing conversations. McCoy et al. (2025) and Le Pennec et al. (2024) demonstrated deep learning's capability to process complex intrapartum signals. However, clinical trust would still be reliant on how obstetric teams respond to alerts under pressure. In the literature, AI seems to be most reliable in the role of a second reader, for early alerts or communication, rather than for obstetric judgment.

Explainability as Risk Communication

Explainability and communication transform risk scores into implementable knowledge. Interpretability tools were represented in several studies, particularly SHAP and feature importance. Khan et al. (2023) applied explainable machine learning to predict risk factors of preterm birth, and Li et al. (2024) applied SHAP to interpret preeclampsia prediction in a Chinese cohort. Hu et al. (2023) and Germaine et al. (2025) also used feature ranking or a history of previous pregnancy as an entry point for more personalised gestational diabetes risk assessment. These are good techniques because they go beyond black-box probability. The review, however, had a systematic discrepancy between technical explainability and human explanation. SHAP plot can inform clinicians that blood pressure, BMI, glucose, parity or previous obstetric history was a risk factor in a score. Still, it does not inform patients about what to do next or the degree of certainty. Human-centred risk communication needs a second step in translation: model feature to clinical meaning, clinical meaning to action plan and action plan to patient choice.

Prediction Performance versus Methodological Weakness

Technically interesting but methodologically weak systems that predict. The main research was evidence-based and demonstrated remarkable performance in several areas. Artzi et al. (2020) predicted gestational diabetes using national EHR data, whereas Kang et al. (2023) added multicentre Korean data to predict time-windowed GDM. Eberhard et al. (2025) reinforced evidence for preeclampsia by predicting and externally validating longitudinally. Le Pennec et al. (2024) enhanced the credibility of fetal monitoring by cross-centre validation of DeepCTG 2.0. Nonetheless, a significant number of studies were retrospective, dependent on datasets, or had limited relevance to clinical outcomes. On 1,611 records, Cubillos et al. (2023) tested a large number of models, making the selection of the model more likely to reward local trends. Zhang et al. (2023) found a significantly greater predictive value for non-preterm than for preterm birth, reflecting how overall accuracy may conceal poor identification of the high-risk group. In the study by Esteban-Escano et al. (2022), a case-control design was employed with a sample of 378 infants, which is not sufficient to confirm proof of concept but cannot address real-time utility. The most trust-relevant studies, according to the methodology, were those that involved external validation, prospective data, explainability or multicentre testing, though these rarely addressed communication outcomes.

Trust Beyond Accuracy: Equity, Privacy and Workflow

Builds trust alongside working outside of AUROC requires equity, privacy and workflow. Figure 3 reveals that equity, privacy, and workflow emerged across the domains; however, they showed reduced intensity compared to validation and explainability. This is the main flaw of the sphere. The majority of model studies considered average performance, whereas women do not care about average care. They are taken care of based on the language, comorbidity, ethnicity, geography, service access and confidence in the health system. The main reason Sufriyana et al. (2020) and Torres-Torres et al. (2023) were significant was that both authors took into account the Indonesian context and a middle-income country, rather than relying mainly on high-income datasets. Artzi et al. (2020) and Kang et al. (2023) were large national or multicentre studies, yet their limitation was population specificity. There was another concern: patient-facing work, where privacy and responsible use of data were paramount. Oprescu et al. (2022) demonstrated that a lack of trust or concerns about data can reduce the desire to use AI.

One of the cross-cutting observations is that the high-risk pregnancy AI is presently performing more of a predictive role than a communicative one. A common conclusion of prediction studies is that clinical implementation would be possible, whereas a common conclusion of studies at the patient level is what they require for implementation. The thing is that there are seldom crossings of these streams. This evidence suggests adopting a hybrid approach: risk algorithms need to be vetted at the local level and subsequently integrated into clinician-led discussions, rather than presented as single alerts. Outputs required by clinicians are actionable and calibrated. Specific, plain-language, emotionally sensitive explanations are required of patients. Audit trails, override mechanisms, and subgroup monitoring should be developed in health systems. It is possible that, without these, AI may add risk labelling while leaving trust unchanged.

Table 1. Primary empirical studies were included in the thematic findings.

Domain	Study	Design and data	Trust or communication relevance	Methodological critique
Patient trust and digital tools	Oprescu et al. (2022)	Prospective qualitative descriptive study with 150 pregnant patients in Spain.	Women wanted responsible, safe and useful AI, as well as clearer explanations.	Strong patient relevance, but preferences were hypothetical and from limited settings.
Patient trust and digital tools	Lazarevic et al. (2023).	Mixed-methods survey and interviews on pregnancy app features.	High interest in communication tools and selected monitoring functions.	Rich user voice, but pandemic context and self-selected digital users limit transferability.
Preeclampsia	Sufriyana et al. (2020).	Nested case-control prediction using Indonesian national insurance	Shows potential for population-scale early warning.	Administrative data offer scale but limited clinical nuance

Domain	Study	Design and data	Trust or communication relevance	Methodological critique
		data.		and communication outcomes.
Preeclampsia	Tarca et al. (2022)	Longitudinal prediction using maternal, biophysical and biochemical markers.	Dynamic risk estimation can support repeated counselling.	A biomarker-dependent strategy may be difficult to implement in lower-resource settings.
Preeclampsia	Torres-Torres et al. (2023)	Machine learning prediction in a middle-income country setting.	Tests transportability beyond high-income datasets.	Resource relevance is strong, but implementation and patient understanding were not tested.
Preeclampsia	Li et al. (2024).	Population-based cohort in China with multiple ML methods and SHAP.	Explainability can support the clinician's interpretation of risk factors.	Single-centre design and limited external validation reduce confidence.
Preeclampsia	Eberhard et al. (2025).	Longitudinal interpretable ML using EHR and external validation.	Personalised repeated prediction supports risk commun	Still needs prospective workflow and patient-facing

Domain	Study	Design and data	Trust or communication relevance	Methodological critique
			ication over time.	explanation studies.
Gestational diabetes	Artzi et al. (2020).	Nationwide Israeli EHR prediction across 588,622 pregnancies.	Early GDM prediction could shift counselling before routine screening.	Excellent scale, but national data may not transfer across ethnic and service contexts.
Gestational diabetes	Hu et al. (2023).	Case-control XGBoost model compared with logistic regression.	Feature ranking may help explain risk contributors.	Case-control design and modest temporal testing limit real-world certainty.
Gestational diabetes	Cubillos et al. (2023).	Twelve ML models using 1,611 pregnancy records.	Low-cost variables could support early intervention in resource-limited care.	A small dataset and many models increase the risk of overfitting.
Gestational diabetes	Kang et al. (2023)	Multicentre Korean retrospective data from 34,387 pregnancies.	Time-windowed predictors can align with antenatal visits.	Ethnic specificity is useful locally but limits global generalisability.
Gestational diabete	Germaine et al.	Retrospective EHR study	Previous pregnancy data	Dependent on EHR

Domain	Study	Design and data	Trust or communication relevance	Methodological critique
s	(2025).	using current and previous pregnancies.	may be used to personalise early risk communication.	completeness and historical data availability.
Preterm birth	Khan et al. (2023)	Explainable ML in a prospective cohort of nulliparous and multiparous women.	SHAP explanations identify risk factors for clinician counselling.	Interpretability was technical; patient comprehension was not measured.
Preterm birth	Goldsztejn & Nehorai (2023).	Deep learning on electrohysterogram recordings from public datasets.	Physiological signals may identify risk beyond what routine history can.	Small and specialised datasets make clinical translation uncertain.
Preterm birth	Zhang et al. (2023).	EHR-based ML prediction study with 5,411 participants.	Routine records may enable earlier identification of PTB risk.	Performance imbalance shows a risk of predicting non-cases better than cases.
Preterm birth	Wang et al. (2025)	Retrospective comparison of logistic regression, random forest, SVM and transformer.	Modern deep learning may improve complex temporal prediction.	Single-institutional data and a retrospective design limit the evidence for workflow.
Fetal monitoring and acidemia	Esteban-Escano et al. (2022).	Case-control EFM study of 378 infants using ML models.	Random forest may support the interpretation of acidemia risk during labour.	A case-control design and a small sample size may overestimate clinical value.
Fetal monitoring and acidemia	Daydulo et al. (2022)	Deep learning from CTU-UHB cardiotocography signals.	Automated fetal distress detection may reduce variability in visual interpretation.	Public dataset modelling lacks testing of live workflow and outcome communication.
Fetal monitoring and acidemia	McCoy et al. (2025).	Deep learning on multisite intrapartum EFM records.	High sensitivity scenarios could support escalation decisions.	Specificity trade-offs may increase intervention unless communication is careful.
Fetal monitoring and acidemia	Le Pennec et al. (2024).	DeepCTG 2.0 multicentre validation for neonatal acidemia.	Cross-centre validation strengthens confidence in the model's robustness.	AUC variation across centres confirms the need for local calibration and audit.

Table 1 summarises the findings from the 20 coded studies. The body of evidence is not evenly distributed. The patient-facing studies directly explored patients' desires for AI and digital monitoring, whereas most prediction studies tested the model rather than communication. Large EHR studies, multicentre fetal-monitoring datasets were the strongest technical evidence. The best source of

trust was the qualitative and mixed-methods studies. The most generally weak aspect of the evidence base was the relationship between prediction and communication: not many of the studies had tested whether women interpreted the risk output, whether it went up or down with the feelings of anxiety, and whether it was better than shared decisions.

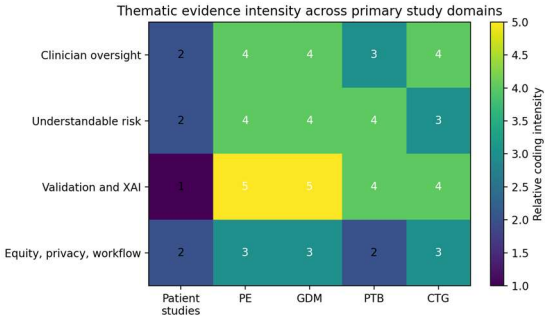


Figure 3. Thematic evidence matrix showing relative coding intensity by evidence domain.

DISCUSSION

The results are consistent with the primary argument of this critique: human-centred trust in AI-assisted high-risk pregnancy care is a relationship-established construct, rather than a model attribute per se. This can be predicted by trust in automation theory, since proper dependence on the system is based on whether users understand the system's purpose, processes and performance (Lee & See, 2004). The primary researches attest to the inefficiency of performance on its own. The technical value of large EHRs like Artzi et al. (2020), Kang et al. (2023) and Eberhard et al. (2025) is evident, but nothing demonstrates that expecting women will comprehend and/or feel safe because of the outlet. On the contrary, Oprescu et al. (2022) and Lazarevic et al. (2023) demonstrate that women can be open to digital tools in case they enhance communication, continuity and responsibility. This contrasts with a model-centred concept of AI and a patient-centred concept of risk communication.

The current results indicate that the perceived usefulness should be reconfigured for high-risk pregnancy rather than the technology acceptance theory. Convenience, speed or access might be usefulness in regular digital health. The protection against avoidable harm, agency preservation, and certainty that professionals are accountable also serve as useful in high-risk pregnancy. An application that tracks swelling can be acceptable if it enables a woman to share a concern with a clinician, as proposed by Lazarevic et al. (2023). An acceptable preeclampsia score can exist when the practitioner clarifies how it will be used to inform the woman about blood pressure care, symptom education and education, rather than merely defining a woman as at risk. Thus, usability is ineffective. The usefulness of the technology involves the communication pathway surrounding the tool.

The results also cast doubt on the future of explainable AI. Theoretically, explainability is the aspect that can instill greater trust in predicting because it exposes the reasons why a model produced a given prediction (Miller, 2019; Rosenbacke et al., 2024). Explainability in the original papers was typically technical, such as SHAP values, explainable feature weights or interpretable longitudinal risk factors (Eberhard et al., 2025; Khan et al., 2023; Li et al., 2024). These are superior to opaque output, yet they may fall short of fostering patient comprehension. The reason is the risk perception theory. A woman might not simply enquire what variables resulted in this score (Slovic, 1987). Thus, the explanation must be emotionally and clinically meaningful. Human-centred AI must deliver technical explanations to clinicians and risk narratives in the form of action to patients.

The most significant variation among primary studies concerns validation. The methodologies of studies such as Le Pennec et al. (2024), McCoy et al. (2025), Kang et al. (2023) and Eberhard et al. (2025) are stronger than those of small single-site or case-control studies, as they employ larger multicentre or external validation designs. Trust is not, however, sealed even with solid validation. Even a fetal-monitoring model with an AUROC above chance might require fine-tuning of the threshold due to potential compromises in favor of specificity and unnecessary actions at high sensitivity. In the same way, the prediction of gestational diabetes at an early stage can help in the lifestyle counselling; they should not be blamed and burdened by the diagnosis in advance. This is why risk communication should specify both benefits and uncertainties. A score is meant to spark discussion about possibilities rather than make a predictive statement.

The least developed dimension is equity. Fairness, accountability and governance were identified in the literature review as a vital component of responsible AI (Char et al., 2018; WHO, 2021). The target populations of the studies varied geographically, including Indonesia, China, Chile, Korea, Israel, Ireland, Spain and multicentre fetal datasets. However subgroup audit was inconsistent. This poses a significant drawback since the result of high-risk pregnancies is socially patterned. Models that have been trained on groups that are better documented might not perform as well on women with less antenatal access, language barriers, migrants, disability, or who are less represented by ethnic groups. When AI-mediated communication provides a woman with confident affirmation when her data

pattern is not well represented, trust is not misplaced. It becomes unsafe. There needs to be a report of calibration, sensitivity, specificity and clinical utility with respect to the relevant subgroups in future research and testing whether the explanations are comprehended across literacy and language groups must occur in the future.

The results can be comparable to the patient-centred care theory. Patient-centred care involves communication, respecting of values and participation in decision-making (Institute of Medicine, 2001). Women found communication tools and responsible design important in this review, with the patient-facing studies shedding more light on these values. The prediction studies had potential clinical usability although they were not typically patient-centred in their measurements. They failed to inquire whether women found the outputs comprehensible, whether clinicians used the outputs in mutual decision-making, or whether alerts affected anxiety and trust. This is not just a reporting gap, but a methodological gap. In case of a human centred trust goal, patient-reported outcomes, decisional conflict, perceived understanding, trust in clinician and system, privacy concern, experience of risk communication should be evaluated.

Another implication is that high-risk pregnancy AI is to be considered as a service intervention. The model is just one aspect. The intervention includes antenatal booking, laboratory testing, the availability of ultrasounds, midwife contact, clinician review, and pathways of escalation, emergency advice, documentation, and follow-up. The predictive value of a model that predicts preeclampsia at 20 weeks cannot be very useful with a care system without the capability to rapidly review or administer aspirin counselling. A cardiotocography algorithm cannot be valuable when it disrupts the working process and does not define whether something should be done. This service viewpoint is the reason why the current results are different to the purely technical review of AI. Technical reviews are required, but it can overstate preparedness when not discussing workflow, communication and governance.

The discussion has thus taken a conservative yet positive stance. High-risk pregnancy risk communication can be enhanced with AI making the risk earlier, more personalised and more reviewable. When it is dark, not well-validated, inequitable or not linked to clinician explanation, it can sabotage trust. Supervised, explainable and accountable decision support is the most defendable approach. Under this model, the clinician is still in charge of interpreting the score; the patient gets an explanation in plain language; performance is audited by the health system; and the tool is altered or discontinued in case of inequitable or hazardous impacts.

Implications for Practice and Research

To be applicable in clinical practice, AI-assisted risk communication ought to be based on five principles. To begin with, each AI-generated risk product must have access to a named clinical action, including repeat blood pressure, glucose, fetal growth check, safety-net and labour-escalation advice. Second, the clinicians need to clarify whether it is a tool to help in screening or the second reader or a monitoring support, as patients must understand that AI does not supplant professional judgement. Third, plain-language and probabilistic explanations must be offered: it is more credible for the tool to indicate that you are at higher-than-average risk, and we will oversee you more closely than if no explanation of your risk is provided. Fourth, health systems are expected to make information on data use and privacy protection available in easily understandable language. Fifth, local teams are expected to oversee false positives, false negatives, interventions, patient anxiety and subgroup performance.

In research, the field must have fewer isolated retrospective models and implementation studies. In the future, primary studies should be carried out by integrating model validation and communication outcomes. Randomised or prospective research would (possibly) determine whether AI-assisted counselling enhances understanding, monitoring compliance, prompt presentation of warning signs, contentment and maternal or neonatal outcomes. The qualitative study ought to comprise women of varying literacy levels, ethnic communities, migration backgrounds and resource backgrounds. Methods are supposed to report on calibration, analysis of decision-curve, missing data, justification of threshold, external validation and subgroup performance. In the absence of these factors, research can be technically interesting but not suitable to high-risk care that requires trust.

A reporting recommendation is also proposed using the thematic analysis. Best practices Obstetric AI research must incorporate a concise communication and governance statement that includes information about who is told about the prediction, who describes the prediction, what happens next, how it is communicated to the patient, how privacy is preserved, how bias is audited, and how clinician override is reported. This statement would render the social pathway of the model observable. It would also assist reviewers to tell the difference between predictive novelty and the implementable care improvement.

CONCLUSION

Human-centred trust on AI-aided risk communication during care of high-risk pregnancy may not be modelable to accuracy. The main evidence explored in this section reveals that AI can be used to predict preeclampsia earlier, gestational diabetes, preterm birth, and fetal compromise, and that expectant women can accept digital assistance when it enhances communication and responsible care. But the evidence has its crucial gaps as well. A variety of models are retrospective, locally validated, technically described and not patient-described, and weakly assessed to equity, privacy and workflow. Autonomous risk labelling is not the most reliable future of obstetric AI. It is patient-centred, clinician-supervised, explainable and locally validated decision support. Risk needs to be more understandable and actionable, rather than more opaque with the help of AI. Trust is developed in high-risk pregnancy in cases where women can realize that the technology is safe, responsible, reasonable and relatable to human care.

REFERENCES

Ahmed, M., Mahmood, S., & Khan, M. (2024). Machine learning models for predicting preeclampsia: A systematic review. BMC Pregnancy and Childbirth, 24, Article 112.

American College of Obstetricians and Gynecologists. (2020). Gestational hypertension and preeclampsia: ACOG Practice Bulletin, Number 222. Obstetrics & Gynecology, 135(6), e237-e260. https://doi.org/10.1097/AOG.0000000000003891

Artzi, N. S., Shilo, S., Hadar, E., Rossman, H., Barbash-Hazan, S., Ben-Haroush, A., Balicer, R. D., Feldman, B., Wiznitzer, A., & Segal, E. (2020). Prediction of gestational diabetes based on nationwide electronic health records. Nature Medicine, 26, 71-76. https://doi.org/10.1038/s41591-019-0724-8

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021). Thematic analysis: A practical guide. SAGE.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983. https://doi.org/10.1056/NEJMp1714229

Correia, V., Mascarenhas, T., & Mascarenhas, M. (2025). Smart pregnancy: AI-driven approaches to personalised maternal and foetal health. Journal of Clinical Medicine, 14(19), 6974. https://doi.org/10.3390/jcm14196974

Cubillos, G., Monckeberg, M., Plaza, A., Morgan, M., Estevez, P. A., Choolani, M., Kemp, M. W., Illanes, S. E., & Perez, C. A. (2023). Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy and Childbirth, 23, 436. https://doi.org/10.1186/s12884-023-05766-4

Davidson, S., Monteith, S., & Glenn, T. (2025). Artificial intelligence in obstetrics: Current applications and future directions. BMJ Medicine, 4, e001032.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340. https://doi.org/10.2307/249008

Daydulo, Y. D., Thamineni, B. L., Dasari, H. K., & Aboye, G. T. (2022). Deep learning based fetal distress detection from time frequency representation of cardiotocogram signal using Morse wavelet: Research study. BMC Medical Informatics and Decision Making, 22, 329. https://doi.org/10.1186/s12911-022-02068-1

Eberhard, B. W., Cohen, R. Y., Wheeler, N., Kleinlein, R., Rigoni, J., Bates, D. W., Gray, K. J., & Kovacheva, V. P. (2025). Development and validation of an interpretable longitudinal preeclampsia risk prediction using machine learning. PLOS ONE, 20(6), e0323873. https://doi.org/10.1371/journal.pone.0323873

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32-64. https://doi.org/10.1518/001872095779049543

Esteban-Escano, J., Castan, B., Castan, S., Choliz-Ezquerro, M., Asensio, C., Laliena, A. R., Sanz-Enguita, G., Sanz, G., Esteban, L. M., & Saviron, R. (2022). Machine learning algorithm to predict acidemia using electronic fetal monitoring recording parameters. Entropy, 24(1), 68. https://doi.org/10.3390/e24010068

Germaine, M., O'Higgins, A. C., Egan, B., & Healy, G. (2025). Evaluation of machine learning models for early prediction of gestational diabetes using retrospective electronic health records from current and previous pregnancies. BMJ Digital Health and AI, 1(1), e000089.

https://doi.org/10.1136/bmjdhai-2025-000089

Goldsztejn, U., & Nehorai, A. (2023). Predicting preterm births from electrohysterogram recordings via deep learning. PLOS ONE, 18(5), e0285219. https://doi.org/10.1371/journal.pone.0285219

Hu, X., Hu, X., Yu, Y., & Wang, J. (2023). Prediction model for gestational diabetes mellitus using the XGBoost machine learning algorithm. Frontiers in Endocrinology, 14, 1105062. https://doi.org/10.3389/fendo.2023.1105062

Institute of Medicine. (2001). Crossing the quality chasm: A new health system for the 21st century. National Academies Press.

International Diabetes Federation. (2025). Gestational diabetes. International Diabetes Federation.

Kang, B. S., Lee, S. U., Hong, S., Choi, S. K., Shin, J. E., Wie, J. H., Jo, Y. S., Kim, Y. H., Kil, K., Chung, Y. H., Jung, K., Hong, H., Park, I. Y., & Ko, H. S. (2023). Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. Scientific Reports, 13, 13156. https://doi.org/10.1038/s41598-023-39680-8

Khan, W., Zaki, N., Ghenimi, N., Ahmad, A., Bian, J., Masud, M. M., Ali, N., Govender, R., & Ahmed, L. A. (2023). Predicting preterm birth using explainable machine learning in a prospective cohort of nulliparous and multiparous pregnant women. PLOS ONE, 18(12), e0293925. https://doi.org/10.1371/journal.pone.0293925

Lazarevic, N., Pizzuti, C., Rosic, G., Boehm, C., Williams, K., & Caillaud, C. (2023). A mixed-methods study exploring women's perceptions and recommendations for a pregnancy app with monitoring tools. npj Digital Medicine, 6, 50. https://doi.org/10.1038/s41746-023-00792-0

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Le Pennec, E., Morin, C., Hoarau, C., et al. (2024). DeepCTG 2.0: Development and validation of a deep learning model to detect neonatal acidemia. Computers in Biology and Medicine, 185, 109448. https://doi.org/10.1016/j.compbiomed.2024.109448

Li, H., Scheinberg, M., Zhao, Z., et al. (2026). Machine learning for dynamic and short-term prediction of preeclampsia using routine clinical data. JAMA Network Open, 9(3), e260359. https://doi.org/10.1001/jamanetworkopen.2026.0359

Li, T., Xu, M., Wang, Y., Wang, Y., Tang, H., Duan, H., Zhao, G., Zheng, M., & Hu, Y. (2024). Prediction model of preeclampsia using machine learning based methods: A population based cohort study in China. Frontiers in Endocrinology, 15, 1345573. https://doi.org/10.3389/fendo.2024.1345573

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335

McCoy, J. A., Levine, L. D., Wan, G., Chivers, C., Teel, J., & La Cava, W. G. (2025). Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. American Journal of Obstetrics and Gynecology, 232(1), 116.e1-116.e9. https://doi.org/10.1016/j.ajog.2024.04.022

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

National Institute for Health and Care Excellence. (2020). Diabetes in pregnancy: Management from preconception to the postnatal period (NICE Guideline NG3).

National Institute for Health and Care Excellence. (2023). Hypertension in pregnancy: Diagnosis and management (NICE Guideline NG133).

Oprescu, A. M., Miro-Amarante, G., Garcia-Diaz, L., Rey, V. E., Chimenea-Toscano, A., Martinez-Martinez, R., & Romero-Ternero, M. C. (2022). Towards a data collection methodology for responsible artificial intelligence in health: A prospective and qualitative study in pregnancy. Information Fusion, 83-84, 53-78. https://doi.org/10.1016/j.inffus.2022.03.011

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Petrozziello, A., Redman, C. W. G., Papageorghiou, A. T., Jordanov, I., Georgieva, A., & Ugwumadu, A. (2019). Computerized cardiotocography analysis for intrapartum fetal monitoring: A systematic review. Acta Obstetricia et Gynecologica Scandinavica, 98(1), 7-15.

Rosenbacke, R., Melhus, A., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. JMIR AI, 3, e53207. https://doi.org/10.2196/53207

Slovic, P. (1987). Perception of risk. Science, 236(4799), 280-285. https://doi.org/10.1126/science.3563507

Sufriyana, H., Wu, Y. W., & Su, E. C. Y. (2020). Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine, 54, 102710. https://doi.org/10.1016/j.ebiom.2020.102710

Tarca, A. L., Taran, A., Romero, R., Jung, E., Paredes, C., Bhatti, G., Ghita, C., Chaiworapongsa, T., Than, N. G., & Hsu, C. D. (2022). Prediction of preeclampsia throughout gestation with maternal characteristics and biophysical and biochemical markers: A longitudinal study. American Journal of Obstetrics and Gynecology, 226(1), 126.e1-126.e22. https://doi.org/10.1016/j.ajog.2021.01.020

Torres-Torres, J., Villafan-Bernal, J. R., Martinez-Portilla, R. J., Hidalgo-Carrera, J. A., Estrada-Gutierrez, G., Adalid-Martinez-Cisneros, R., Rojas-Zepeda, L., & Espino-y-Sosa, S. (2023). Performance of a machine-learning approach for prediction of pre-eclampsia in a middle-income country. Ultrasound in Obstetrics & Gynecology, 63(3), 350-357. https://doi.org/10.1002/uog.27510

UNICEF. (2025). Trends in maternal mortality 2000 to 2023. UNICEF Data.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36(1), 157-178.

Wang, H., Li, N., Chivese, T., Werfalli, M., Sun, H., Yuen, L., Hoegfeldt, C. A., Powe, C. E., Immanuel, J., Karuranga, S., Divakar, H., Levitt, N., Li, C., Simmons, D., & Yang, X. (2021). IDF diabetes atlas: Estimation of global and regional gestational diabetes mellitus prevalence for 2021 by International Association of Diabetes in Pregnancy Study Group's criteria. Diabetes Research and Clinical Practice, 183, 109050. https://doi.org/10.1016/j.diabres.2021.109050

Wang, Y., Li, J., Pan, Y., et al. (2025). Deep learning in predicting preterm birth: A comparative study of algorithms on real-world obstetric data. Maternal-Fetal Medicine, 7(2), 88-97. https://doi.org/10.1097/FM9.0000000000000236

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization.

World Health Organization. (2023). Preterm birth. World Health Organization.

World Health Organization. (2025a). Maternal mortality. World Health Organization.

World Health Organization. (2025b). Pre-eclampsia. World Health Organization.

Young, A. T., Amara, D., Bhattacharya, A., & Wei, M. L. (2021). Patient and general public attitudes towards clinical artificial intelligence: A mixed methods systematic review. The Lancet Digital Health, 3(9), e599-e611. https://doi.org/10.1016/S2589-7500(21)00132-1

Zhang, Y., Du, S., Hu, T., Xu, S., Lu, H., Xu, C., Li, J., & Zhu, X. (2023). Establishment of a model for predicting preterm birth based on the machine learning algorithm. BMC Pregnancy and Childbirth, 23, 735. https://doi.org/10.1186/s12884-023-06058-7