DIGITAL HEALTH AND UHC
TECHNICAL WORKING PAPER SERIES · No. 33

ADHERE+: development, training and validation of the intrapartum risk-stratification model

An on-device gradient-boosted clinical decision-support model for maternal care in Ethiopian primary health care

Sileshi Amare, MD, MPH · **Muluken Getahun, PhD**

Bahir Dar University, Bahir Dar, Ethiopia

Working paper — pre-publication
February 2026 · Model version adhere-eth-2.0
Bahir Dar University

About this paper

This working paper is a product of Bahir Dar University. The ADHERE+ platform was developed by Bahir Dar University's Faculty of Computing, with its School of Public Health. EPIC Health Systems collaborated with Bahir Dar University on the pilot evaluation of the ADHERE+ tool.

How to cite

Amare S, Getahun M. ADHERE+: development, training and validation of the intrapartum risk-stratification model. Digital Health and UHC Technical Working Paper Series No. 33. Bahir Dar, Ethiopia: Bahir Dar University; 2026.

Disclaimer

Working papers in this series are circulated to encourage discussion and comment. They have not been peer reviewed. The findings, interpretations and conclusions are those of the authors and do not necessarily represent the views of Bahir Dar University or of collaborating institutions.

© 2026 Bahir Dar University. Reproduction is permitted with attribution.

Abstract

Background. Ethiopia's maternal mortality ratio remains approximately 401 per 100,000[1] live births, with the Amhara Region persistently above the national average, and partograph utilisation below 30 % in many primary healthcare (PHC) facilities. Timely recognition of intrapartum deterioration is constrained by workforce shortages, inconsistent documentation, and the cognitive burden of synthesising numerous clinical parameters under time pressure. The ADHERE+ platform extends a previously published, rule-based electronic partograph and Safe Childbirth Checklist[2] tool with a machine-learning risk-stratification layer intended to assist, not replace, clinical judgement at the point of care.

Methods. We describe the development, training, and validation of adhere-eth-2.0, a gradient-boosted decision-tree model that estimates the probability of intrapartum complication at each partograph examination and returns a green/amber/red stratification. The model consumes 22 routinely collected intrapartum, historical, and vital-sign features. It was developed and validated on a validation dataset of 12,824 partograph examination-records drawn from 2,852 labour episodes across 12 facilities. Examination-level outcome prevalence was 25.2 %. To avoid leakage from repeated examinations of the same labour, the dataset was split by grouping all examinations belonging to a given labour episode into either the training or the test partition (70/30 hold-out: 8,983 training exams, 3,841 test exams), and robustness was further assessed by grouped five-fold cross-validation. The exported model was also implemented as an on-device JavaScript evaluator for offline scoring, and its numerical agreement with the reference implementation was checked directly.

Results. On the grouped hold-out set, the model achieved an area under the receiver operating characteristic curve (AUROC) of 0.809 and a Brier score[3] of 0.141. Grouped five-fold cross-validation gave a closely consistent mean AUROC of 0.809 (standard deviation 0.005; fold range 0.801-0.816), indicating stable discrimination across resampling. At a pre-specified amber operating threshold of 0.33, sensitivity was 0.62, specificity 0.83, positive predictive value 0.57, and negative predictive value 0.86; at a red threshold of 0.60, sensitivity was 0.36, specificity 0.96, positive predictive value 0.75, and negative predictive value 0.81. Discrimination was reasonably stable across clinically relevant subgroups, with AUROC ranging from 0.77 in women with a prior caesarean section to 0.86 in grand-multiparous women. Calibration across probability deciles tracked the identity line reasonably closely at low-to-moderate predicted risk, with expected divergence at sparser high-risk deciles.

Conclusions. These results demonstrate that a compact, interpretable gradient-boosted model can achieve clinically plausible discrimination and calibration on a rigorously leakage-controlled, grouped evaluation, and that it can be deployed for offline point-of-care scoring without loss of numerical fidelity.

---

[1]Central Statistical Agency (Ethiopia), ICF. Ethiopia Demographic and Health Survey 2016. Addis Ababa, Ethiopia, and Rockville, Maryland, USA: CSA and ICF; 2017 (pregnancy-related mortality ratio 412 per 100,000 live births).

[2]World Health Organization. WHO Safe Childbirth Checklist implementation guide: improving the quality of facility-based delivery for mothers and newborns. Geneva: World Health Organization; 2015.

[3]Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138.

Keywords: intrapartum risk stratification; partograph; machine learning; gradient-boosted decision trees[4]; clinical decision support; low-resource settings; Ethiopia; maternal health; on-device inference; model validation.

1. Introduction

1.1 The maternal health challenge in Ethiopian primary care

Maternal mortality remains one of the most persistent indicators of health-system performance in low- and middle-income countries, and Ethiopia continues to report a maternal mortality ratio of approximately 401 per 100,000 live births, a figure that masks substantial sub-national variation. The Amhara Region, where the present work is situated, has consistently been reported above the national average, reflecting a combination of geographic dispersion of population, variable road and referral infrastructure, and heterogeneous capacity across the primary healthcare tier. The great majority of intrapartum deaths and near-miss[5] events in such settings are not the result of a single catastrophic and unforeseeable event, but of a recognised complication - obstructed or prolonged labour, hypertensive disease of pregnancy, intrapartum fetal distress, maternal sepsis, or antepartum haemorrhage - that was not identified, or not acted upon, in time. This is sometimes described in the health-systems literature as the problem of the 'third delay': delay in receiving adequate care once a woman has reached a facility, following delays in deciding to seek care and in reaching a facility at all. Reducing the third delay is, in large part, a problem of clinical information: ensuring that the right observation is made, recorded, and correctly interpreted, and that the correct escalation occurs without unnecessary loss of time.

The partograph, endorsed by the World Health Organization as the principal tool for monitoring labour progress and maternal-fetal wellbeing, was designed precisely to support this task: a single, structured chart on which cervical dilation, descent, fetal heart rate, contraction frequency, and maternal vital signs are plotted against time, so that deviation from expected progress becomes visually apparent before it becomes clinically catastrophic. Despite decades of promotion, however, partograph utilisation remains below 30 % in many Ethiopian PHC facilities, a finding replicated across numerous sub-Saharan African health systems. The reasons are well documented and largely structural rather than attitudinal: heavy caseloads relative to available skilled birth attendants, inconsistent supply of pre-printed charts, limited in-service mentorship and supervision, high staff turnover eroding institutional practice, and the sheer cognitive load of manually plotting, and then correctly interpreting, a dozen or more parameters across a labour that may last many hours. Where the partograph is used, it is frequently completed retrospectively or partially, undermining its function as an immediate early-warning instrument. The consequence is a persistent gap between the theoretical promise of structured labour monitoring and its practical delivery at the point of care, a gap that manifests as delayed recognition of, and delayed response to, treatable intrapartum deterioration.

---

[4]Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001;29(5):1189-1232.

[5]Prevalence and determinants of maternal near miss in Ethiopia: a systematic review and meta-analysis. Systematic Reviews. 2023. doi:10.1186/s13643-025-02770-x.

These workforce and systems constraints are not peculiar to Ethiopia, but they are especially acute in rural and PHC-tier facilities, where a single mid-level provider - often a midwife or health officer without continuous physician back-up - may be simultaneously responsible for several labouring women, antenatal consultations, and other primary-care duties. Under such conditions, any tool intended to improve intrapartum safety must not add net documentation burden; it must instead make existing, already-collected observations more informative, more visible, and more actionable, without requiring additional data collection, additional infrastructure, or reliable connectivity. This constraint - improve decision quality without increasing burden - is the organising design principle behind the ADHERE+ platform described in this paper, and it motivates both the choice of the partograph as the substrate for augmentation and the choice of an offline, on-device computational approach described in Section 1.3.

1.2 The partograph and the case for point-of-care risk stratification

The partograph's clinical logic rests on comparing observed values - most centrally the rate of cervical dilation - against pre-defined alert and action lines, so that a provider can identify labour that is progressing more slowly than expected and intervene before obstruction or exhaustion supervenes. This is, in effect, a simple univariate or bivariate decision rule applied to a single parameter (cervical dilation against time), supplemented by separate, largely independent thresholds for fetal heart rate, moulding, and maternal vital signs. Each of these thresholds is clinically sound in isolation, and the alert/action cervicograph remains, and should remain, the backbone of intrapartum monitoring. A woman with borderline-slow cervical dilation, mild moulding, a marginally raised pulse, and a long interval since rupture of membranes may not trigger any single alert line, yet the joint pattern may be more informative than any one observation taken alone.

This observation motivates the case for point-of-care risk stratification: a computational layer that takes the same observations a provider already records on the partograph - cervical dilation and its rate of change, fetal heart rate, contraction frequency, moulding, meconium grade, maternal vital signs, obstetric history, and gestational age - and combines them into a single, continuously updated estimate of the probability of an adverse intrapartum outcome, expressed as an intuitive green/amber/red band rather than as a raw probability. Such a layer does not replace the alert/action lines; it sits alongside them, surfacing a synthesis of multiple weak signals that might individually escape attention but that, combined, indicate a meaningfully elevated risk. Because the underlying observations are already part of routine practice, the marginal data-collection burden on the provider is close to zero: the risk estimate is a re-use of existing information rather than a demand for new information. This is a materially different proposition from many digital-health interventions that require additional structured data entry, and it is central to the platform's plausibility as a low-burden PHC tool.

A further argument for risk stratification, rather than reliance on fixed alert lines alone, concerns the heterogeneity of risk across the obstetric population. A given cervical dilation rate may carry different prognostic weight depending on parity, prior caesarean section, gestational age, and the presence of other risk factors; a purely univariate threshold cannot easily encode these interactions without becoming an unwieldy set of conditional rules. A statistical model that is trained across a large number of

examinations can, in principle, learn these interactions directly from the data and express them as a single composite score, provided that the model is transparent about which findings are driving a given estimate - a requirement addressed directly in the explainability design described in Section 1.4 and revisited in the discussion of clinical safety later in this paper.

1.3 Why on-device machine learning for low-resource primary healthcare

Having established the clinical case for a composite, multivariate risk estimate, the choice of implementation strategy is itself consequential. Many clinical machine-learning systems are designed around continuous connectivity to a central server: observations are transmitted, scored remotely, and a result returned. This architecture is poorly matched to the operating conditions of Ethiopian PHC facilities, where electricity supply may be intermittent, mobile data connectivity may be weak, absent, or costly, and the computing hardware available at the facility level is typically a low-cost Android device rather than a networked workstation. A system that depends on connectivity to function is a system that will, at the moments of greatest clinical urgency, most often during power outages, storms, or network congestion, silently fail exactly when it is most needed.

The ADHERE+ platform therefore adopts an offline-first design in which the risk model itself, exported as a compact JSON representation of the underlying decision trees, is scored entirely on-device by a JavaScript evaluator running within the browser or wrapped native application, with no dependency on network connectivity for the act of scoring. Data entered by the provider is queued locally and synchronised opportunistically when connectivity becomes available, using idempotency keys so that a queued action is applied exactly once even if synchronisation is retried. This design choice has three further, related justifications beyond simple availability. First, it is inexpensive to run: gradient-boosted decision trees of the modest size used here (240 trees, maximum depth 3) impose negligible computational demand on even low-specification Android hardware, in contrast to larger neural architectures that may require dedicated accelerators or substantial memory. Second, it supports transparency: because the model is a finite, inspectable set of decision trees rather than an opaque deep network, the specific features contributing to a given prediction can be extracted and displayed to the provider directly from the model's internal structure, rather than approximated post-hoc. Third, it is consistent with a human-in-the-loop safety philosophy: the on-device estimate is presented as one input among several - alongside the deterministic red-flag danger-sign layer and the transparent, aggregate-weighted Modified Early Obstetric Warning Score - and every alert requires the provider to acknowledge it and record an independent clinical assessment, with any override logged together with a reason code. The model is designed, throughout, to prompt and inform clinical judgement, not to substitute for it.

Gradient-boosted decision trees were selected over alternative model families for reasons developed further in Section 2.1, but it is worth noting here, in the context of the low-resource deployment target, that this model family offers a favourable combination of predictive performance on structured tabular data, small exported model size, fast evaluation, and straightforward interpretability - all properties that matter disproportionately when the target runtime is an offline, low-cost mobile device operated by a busy clinician rather than a data centre operated by a specialist analytics team.

1.4 From ADHERE to ADHERE+

It is important, for scientific clarity, to distinguish two separate bodies of evidence. The first is the original ADHERE system: a deterministic, rule-based electronic partograph and electronic WHO Safe Childbirth Checklist, incorporating danger-sign prompts and suggested actions but containing no machine-learning component. This tool was evaluated in a published, peer-reviewed quasi-experimental difference-in-differences study (Nigatu et al., *Reproductive Health*, 2026), conducted across three hospitals and six health centres (five intervention, four control) over a twelve-month implementation window from June 2022 to May 2023, drawing on 2,190 labour-and-delivery charts. That study found that the ADHERE intervention increased the partograph adherence score (0-9 scale) by 4.27 units[6] (95 % CI 1.80 to 5.09) relative to control, and increased WHO Safe Childbirth Checklist completion by 9.3 percentage points on admission (95 % CI 3.6 to 15.1), 22.3 percentage points before birth (95 % CI 12.1 to 32.5), and 15.2 percentage points after birth (95 % CI 7.2 to 23.1), with no significant effect before discharge.

The second body of evidence, and the subject of this working paper, concerns ADHERE+: a subsequent enhancement that adds a machine-learning risk-stratification layer, the deterministic national-protocol expert system, Bayesian longitudinal risk tracking, and DHIS2[7]/FHIR[8]-compatible reporting analytics to the original rule-based platform. The machine-learning layer described here was developed and validated after the published implementation window described above; it was not present in, and did not contribute to, the results of the Nigatu et al. (2026) evaluation. This paper is exclusively concerned with the second body of evidence: the development, training, and validation of the ADHERE+ maternal intrapartum risk-stratification model, together with a companion newborn resuscitation-need model presented as a proof-of-concept enhancement.

1.5 Objectives and contributions of this paper

This working paper has four specific objectives. First, to document, in sufficient methodological detail for independent scrutiny and reproduction, the development of the adhere-eth-2.0 maternal intrapartum risk-stratification model: its input features, its architecture, its training procedure, and the composition of the validation dataset on which it was built. Second, to report the model's validation performance using a leakage-controlled, grouped evaluation design appropriate to serial, repeated-measures intrapartum data, including discrimination, calibration, decision-relevant operating points, and subgroup performance across clinically important strata. Third, to describe the on-device implementation of the model and to report the numerical fidelity of that implementation relative to the reference model, since a risk-stratification tool that behaves differently offline than in its original training environment would undermine the safety case for its deployment.

---

[6]Nigatu D, Azage M, Misgan E, Enquobahrie DA, Kebebaw T, Abate E, Alemneh E, Woldie M, Girma T. Effect of an eHealth intervention (ADHERE) on adherence to the partograph and WHO Safe Childbirth Checklist in intrapartum care: implementation research in Ethiopia. Reproductive Health. 2026;23:59. doi:10.1186/s12978-026-02279-z.

[7]Dehnavieh R, Haghdoost A, Khosravi A, et al. The District Health Information System (DHIS2): a literature review and meta-synthesis of its strengths and operational challenges. Health Information Management Journal. 2019;48(2):62-75.

[8]HL7 International. HL7 FHIR Release 4 (Fast Healthcare Interoperability Resources). Health Level Seven International; 2019. Available at: hl7.org/fhir.

2. Background and related work

2.1 Machine learning for intrapartum and obstetric risk prediction

The application of statistical and machine-learning methods to obstetric risk prediction has grown substantially over the past decade, spanning tasks as varied as antenatal prediction of pre-eclampsia, prediction of caesarean-section likelihood from admission features, cardiotocography interpretation, and, most closely related to the present work, intrapartum prediction of fetal distress, obstructed labour, and composite adverse-outcome risk from serially recorded labour observations. Across this literature, a recurring and important finding is that structured, tabular clinical data of the kind recorded on a partograph - continuous physiological measurements, ordinal clinical grades, and binary risk flags, typically numbering in the tens of features - is generally better suited to tree-based ensemble methods than to the deep neural architectures that dominate image- and text-oriented machine learning. This is partly a consequence of dataset scale, since deep networks generally require substantially larger training sets than are available in most single- or multi-facility obstetric cohorts to avoid overfitting, and partly a consequence of the structure of the data itself: tabular clinical features often exhibit threshold effects, non-monotonic relationships, and interactions between a moderate number of variables, all of which decision-tree ensembles capture naturally through recursive partitioning, whereas they must be learned less directly by architectures designed around continuous, differentiable feature spaces.

Gradient boosting, in which successive shallow decision trees are fitted to the residual error of the ensemble built so far, has become a particularly strong default choice for structured clinical prediction tasks of this kind, consistently performing competitively with, or better than, more complex alternatives in benchmark comparisons on tabular healthcare data, while retaining several practical advantages that matter for a deployed clinical tool. Shallow trees (in the present model, a maximum depth of three) are individually simple and nearly human-readable, so that the path taken through any single tree for a given examination can be traced and explained. The ensemble as a whole remains compact - here, 240 trees at a learning rate of 0.06, trained with row subsampling of 0.9 to reduce variance and guard against overfitting to any particular subset of examinations - and can be exported as a small, static data structure (a JSON description of split points and leaf values) that is trivial to evaluate on constrained hardware without specialised libraries, a property exploited directly in the on-device scoring architecture described in Section 1.3. These properties - strong tabular performance, interpretability at the level of individual trees, and lightweight deployment - together motivate the choice of gradient-boosted decision trees as the modelling approach for both the maternal intrapartum risk model and the companion newborn resuscitation-need model.

It is also relevant to situate this work against the broader, and larger, literature applying machine learning to intrapartum fetal heart rate (cardiotocography) interpretation specifically, which has attracted particular research attention because of the well-documented inter- and intra-observer variability in visual cardiotocograph interpretation among clinicians. The ADHERE+ maternal model differs from most of that literature in scope: rather than attempting to interpret a continuous physiological waveform, it operates

on the same discrete, periodically recorded observations already captured on a paper or electronic partograph, an approach that trades the potentially richer signal available in continuous monitoring for compatibility with the intermittent, resource-constrained observation practices that actually obtain in most PHC-tier facilities in Ethiopia and comparable settings, where continuous electronic fetal monitoring is rarely available.

2.2 Evaluation methodology for clinical prediction models

Rigorous evaluation of a clinical prediction model requires attention to several distinct properties, each of which answers a different question and none of which, alone, is sufficient. Discrimination - most commonly summarised by the area under the receiver operating characteristic curve (AUROC) - measures the model's ability to rank a case that will experience the outcome above a case that will not, but it says nothing about whether the model's stated probabilities are themselves trustworthy in absolute terms. Calibration addresses precisely that latter question: it asks whether, among all examinations to which the model assigns a predicted probability of, say, 0.35, the observed proportion of adverse outcomes is indeed close to 35 %. For this reason, this paper reports calibration across probability deciles alongside discrimination, rather than discrimination alone.

Beyond discrimination and calibration, a third and increasingly emphasised property is clinical utility, typically assessed through decision-curve analysis, which asks whether using the model to guide a binary decision (for example, escalate versus do not escalate) yields a greater net benefit than simple default strategies of treating everyone or treating no one, across a plausible range of decision thresholds. Decision-curve and net-benefit analyses are particularly important for a safety tool of this kind because discrimination and calibration statistics, while necessary, do not by themselves establish that acting on the model's output would be preferable to the available default strategies at the thresholds a provider would actually use; that judgement depends on the relative costs of false alarms versus missed escalations at the specific operating points a deployment intends to use, which is precisely why this paper reports concrete sensitivity, specificity, and predictive-value figures at the amber and red thresholds actually implemented in the platform, rather than discrimination statistics alone.

A methodological hazard specific to intrapartum data of this kind deserves particular emphasis: data leakage arising from serial, repeated-measures structure. A partograph generates multiple examination-records over the course of a single labour, and successive examinations from the same labour are highly correlated, sharing the same underlying maternal history, obstetric trajectory, and much of the same physiological trend. If examinations are split into training and test sets independently of the labour episode to which they belong, it becomes possible - indeed likely - for examinations from the same labour to appear in both partitions, so that the model is effectively tested on data drawn from a case it has already, in part, seen during training. This inflates apparent performance and produces an optimistic, non-generalisable estimate of discrimination and calibration. The methodologically correct approach, and the one adopted for the maternal and newborn models described in this paper, is to split at the level of the labour episode ('grouped' splitting), so that every examination belonging to a given labour is assigned entirely to either the training partition or the test partition, never both. This is why the present evaluation

reports a 70/30 hold-out split of 8,983 training examinations against 3,841 test examinations, grouped by labour, and supplements the single hold-out estimate with grouped five-fold cross-validation to characterise the stability of the discrimination estimate across resampling, rather than relying on a single train/test split that could, by chance, be favourable or unfavourable.

2.3 On-device and offline clinical decision support in low- and middle-income settings

A growing literature on digital health in low- and middle-income countries has documented the practical necessity of designing for intermittent connectivity and constrained hardware, rather than treating offline operation as an edge case to be tolerated. Studies of mobile health (mHealth) tool adoption across sub-Saharan Africa have repeatedly identified connectivity gaps, unreliable electricity supply, and the cost of mobile data as material barriers to sustained use of cloud-dependent clinical applications, particularly at the primary healthcare tier where facilities may lack even basic institutional Wi-Fi. Offline-first architectures - in which the application functions fully without a network connection and synchronises opportunistically when connectivity is available - have accordingly been identified as a design requirement, rather than a preference, for digital clinical tools intended for sustained PHC-tier use in such settings. This is the architectural tradition within which the ADHERE+ platform's offline-first progressive web application, service-worker caching strategy, and on-device model scoring are situated.

On-device inference for structured clinical risk models - as distinct from on-device inference for large image or language models, which typically remains infeasible on low-cost Android hardware - is comparatively tractable precisely because gradient-boosted tree ensembles of modest size, of the kind used here, impose minimal memory and computational requirements. Evaluating such a model requires only traversing a bounded number of shallow decision trees and summing leaf contributions before a final logistic transform, an operation well within the capability of essentially any modern smartphone processor, including low-cost devices that form the realistic hardware baseline in many PHC facilities in Ethiopia and comparable settings. This tractability is what allows the platform to commit to full offline scoring, rather than a degraded offline mode with reduced functionality, and it is a material point of difference from clinical machine-learning deployments elsewhere that assume continuous connectivity to a hosted inference service.

A further consideration specific to low-resource deployment concerns validation of the deployed artefact itself, not merely the model in the abstract. Because the on-device JavaScript evaluator is a separate software implementation from the Python environment in which the model was trained, it is necessary to confirm directly that the exported model, as it will actually run in the field, reproduces the reference implementation's output with acceptable numerical fidelity; a discrepancy here would mean that the performance figures reported for the trained model do not actually describe the behaviour of the deployed tool. This paper reports that check explicitly (on-device parity to within approximately 1 x 10^-5), reflecting a broader principle that offline clinical decision-support tools for low-resource settings should validate the deployed implementation, not only the model that was trained, as part of establishing that reported performance figures apply to what clinicians will actually encounter on-device.

3. Methods

3.1 Overview and reporting standards

This section describes the methods used to develop, export and validate the ADHERE+ maternal intrapartum risk-stratification model, version adhere-eth-2.0, which forms the machine-learning e-partograph module of the ADHERE+ decision-support platform. The account is structured to follow, as closely as the current stage of the work allows, the reporting logic of the TRIPOD statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis), which is the field-standard checklist for describing the development and validation of clinical prediction models. We adopt TRIPOD's structure — cohort description, outcome and predictor definitions, model-development procedure, and validation design and metrics — because it provides a discipline that is useful irrespective of the setting in which a model is developed: it forces explicit statements of how examination-records were structured, how the outcome was defined, how missing data were handled, how the model was specified and fitted, and how its performance was estimated and reported.

3.2 Validation dataset structure

The validation dataset comprises 12,824 partograph examination-records nested within 2,852 distinct labour episodes, distributed across 12 facilities. Each labour episode contributes multiple examination-rows, corresponding to the successive partograph assessments taken over the course of that labour, so that the examination-record — rather than the labour episode — is the fundamental unit of analysis for model fitting and evaluation throughout this paper.

Working with a dataset whose fundamental unit is the examination-record, rather than the labour episode, has direct methodological consequences that recur throughout this section: because multiple examination-rows drawn from the same labour necessarily share woman-level attributes and overlapping early-labour history, every stage of model development and validation described below — feature construction, the definition of the outcome, and, most consequentially, the design of the train/test split — must treat the labour episode, rather than the individual examination, as the unit that determines independence between rows. Section 3.5 sets out how the outcome is defined at examination level; Section 3.8 sets out how this labour-level dependency is handled in the validation design.

3.4 Input features

The maternal intrapartum risk model uses 22 input features drawn from the partograph examination record, the woman's registration record, and the recorded amniotic-fluid finding. Table 1 presents the full data dictionary: each feature's short field name, a plain-language description, and its unit or permitted range. Three features are sourced outside the examination row itself and are worth flagging explicitly. Maternal age, parity and gestational age are read from the woman's registration record rather than being re-entered at each partograph examination, with neutral fallback values applied where the registration record leaves a field blank, so that a missing registration value does not silently propagate as a spurious zero or default risk-bearing value. The meconium / amniotic-fluid-grade feature is read from the recorded

amniotic-fluid finding taken at vaginal examination, rather than from a separate free-standing field, consistent with how liquor colour is actually documented on a paper or electronic partograph.

Table 1. Data dictionary — 22 input features, maternal intrapartum risk model (adhere-eth-2.0).

Field	Description	Unit / range
hrs	Hours elapsed in active labour at the time of examination	hours
cvx	Cervical dilatation at examination	cm (monotonic within labour)
cvx_rate	Cervical dilatation rate, derived from successive dilatation readings	cm/hour
fhr	Fetal heart rate at examination	beats per minute
ctx	Contraction frequency	contractions per 10 minutes
mld	Moulding of the fetal skull	0–3 (grade)
meconium	Amniotic-fluid / meconium grade; read from the recorded amniotic-fluid finding	0–3 (grade); structurally missing before membrane rupture
sbp	Maternal systolic blood pressure	mmHg
dbp	Maternal diastolic blood pressure	mmHg
urine_prot	Urine protein	0–3 (grade)
temp	Maternal temperature	°C
pulse	Maternal pulse rate	beats per minute
bleeding	Vaginal bleeding present	0/1 (binary)
headache	Severe headache present	0/1 (binary)
blurred	Blurred vision present	0/1 (binary)
epigastric	Epigastric pain present	0/1 (binary)
clonus	Clonus present	0/1 (binary)
age	Maternal age; read from the woman's registration record with neutral fallback	years
parity	Parity; read from the woman's registration record with neutral fallback	count
ga	Gestational age; read from the woman's registration record with neutral fallback	weeks
prior_cs	Prior caesarean section	0/1 (binary)
rom_hours	Hours elapsed since rupture of membranes	hours

All 22 features are constructed to be available, in principle, at the point of any partograph examination during active labour, so that the model can be scored immediately as each new examination is entered, rather than requiring information that would only become available retrospectively (such as final mode of delivery or birth outcome). Categorical and grade-type features (moulding, meconium, urine protein) are represented as small ordinal integer scales matching the grading conventions already used on the paper partograph, so that no additional coding burden is placed on the provider entering data at the point of care. Continuous features (vital signs, cervical dilatation and its derived rate, hours in labour, hours since rupture of membranes) retain their natural clinical units throughout, rather than being pre-transformed, so that the split thresholds learned by the model, and any post-hoc inspection of feature importance, remain directly interpretable by a clinical reader.

3.5 Outcome definition

The modelled outcome is a binary intrapartum adverse-outcome label, defined at the level of each individual partograph examination rather than once per labour. At each examination, the label takes the value 1 if the labour is, at or by that point, associated with one or more of five modelled intrapartum complications — obstructed or prolonged labour, pre-eclampsia or eclampsia, intrapartum fetal distress, maternal or intrapartum sepsis, or antepartum haemorrhage — and 0 otherwise. Defining the outcome at the examination level, rather than once per labour, means that a single labour episode contributes multiple labelled rows across its recorded examinations, and that the label can, and typically does, evolve across a labour's own examination sequence as the clinical picture develops — a labour need not display its complication from its first recorded examination onward. Across the full validation dataset of 12,824 examination-records, the examination-level prevalence of the adverse-outcome label is 25.2%. This single figure aggregates across all five component complications and across all stages of labour represented in the dataset; it is not a claim about the prevalence of any one complication in isolation.

Framing the outcome at examination level, rather than at labour level, was a deliberate design choice intended to match the way the model is actually used at the point of care: a provider scoring the tool wants an estimate of risk at the current examination, given everything recorded up to and including that examination, not a single static risk estimate assigned to the labour as a whole at admission. This examination-level framing does, however, introduce a methodological complexity that is addressed directly in the validation design (Section 3.8): because multiple examination-rows from the same labour share underlying woman-level and early-labour information, a naïve random split of examination-rows into training and test sets would allow rows from the same labour to appear on both sides of the split, producing a form of information leakage that would inflate apparent performance. The grouped-by-labour splitting strategy described below exists specifically to prevent this.

3.6 Model architecture

The maternal intrapartum risk model is a gradient-boosted decision-tree ensemble, comprising 240 trees, each constrained to a maximum depth of 3, fitted with a learning rate of 0.06 and a row-subsample fraction of 0.9 at each boosting iteration. Gradient boosting builds the ensemble sequentially: each new tree is fitted to the residual gradient of the loss function with respect to the current ensemble's predictions, so that each successive tree corrects, in a small and regularised increment (governed by the learning rate), the errors left by the trees fitted before it. Constraining each tree to a maximum depth of 3 restricts every individual tree to a small number of sequential feature splits, which limits the complexity, and hence the overfitting capacity, of any single tree while still allowing the ensemble as a whole, across 240 trees, to represent non-linear interactions between features (for example, the joint effect of a rising fetal heart rate and a prolonged rupture-of-membranes interval). Subsampling 90% of rows at each boosting iteration introduces an additional stochastic regularisation, in the manner of stochastic gradient boosting, which tends to improve generalisation and reduce variance in the fitted ensemble relative to fitting every tree on the identical full training set.

Gradient-boosted decision trees were selected in preference to a deep-learning architecture (for example, a multilayer perceptron or a recurrent or attention-based sequence model) for four reasons specific to

this application. First, on tabular, structured clinical data of the kind a partograph produces — a modest number of well-defined, individually meaningful features, rather than raw image, text or waveform input — gradient-boosted tree ensembles have repeatedly been shown in the wider machine-learning literature to match or exceed the predictive performance of deep neural architectures, while requiring substantially less training data to reach that performance; this matters directly given that the validation dataset, at 12,824 examination-records, is modest by deep-learning standards. Second, tree ensembles offer materially better interpretability: the contribution of each feature to a given prediction can be read directly from the split structure of the constituent trees, which is a precondition for the explainability requirement described in the companion working paper on clinical safety (each risk estimate must be accompanied by the specific contributing findings that drove it) and would be considerably harder to deliver transparently from a deep neural network. Third, gradient-boosted trees are computationally inexpensive to train and to score, which matters both for the practicality of iterative model development and, more importantly, for the deployment requirement described in Section 3.7 below. Fourth, and following directly from the third point, a shallow, modestly sized tree ensemble exports to a compact representation that can be evaluated on a low-cost Android handset with no network connectivity, whereas a deep neural network of comparable predictive capacity would typically require a materially larger exported artefact and a more demanding runtime, working against the offline-first, low-cost-device deployment model that is central to the ADHERE+ platform's design (see the companion working paper on system architecture).

3.7 On-device deployment

The fitted gradient-boosted ensemble is exported to a compact JSON tree representation: each of the 240 trees is serialised as a small nested structure recording, at each internal node, the splitting feature and threshold, and, at each leaf, the contribution that leaf makes to the model's additive log-odds output. This JSON export is bundled into the ADHERE+ progressive web application and cached by the platform's service worker (cache-first for models and JSON, as described in the companion architecture working paper), so that the model is available for scoring entirely offline, with no server round-trip required at the point of care.

A purpose-written JavaScript evaluator, executed on-device, walks each of the 240 exported trees for a given set of input feature values, following the split conditions recorded at each internal node down to the corresponding leaf, and sums the leaf contributions across all 240 trees to obtain an additive log-odds score. This log-odds score is then passed through the logistic transform to yield a predicted probability on the conventional [0, 1] risk scale that is presented to the provider. Any input feature that is missing at the point of scoring — because a field was left blank, not yet assessed, or not yet clinically applicable (for example, meconium grade before membrane rupture) — is imputed on-device using a stored median-default value for that feature, calculated from, and matching, the same feature's median in the training partition used to fit the reference model; this ensures that the on-device evaluator's handling of missing inputs is consistent with, rather than merely similar to, the assumptions embedded in the trained ensemble. On-device parity testing confirms that the JavaScript evaluator reproduces the reference model's output to within approximately 1e-5, which we treat as the acceptable numerical tolerance for the exported artefact.

The resulting predicted probability is mapped to one of three colour-coded risk bands — green, amber, or red — using two fixed decision thresholds: an amber threshold of 0.33 and a red threshold of 0.60. A predicted probability below 0.33 is displayed as green (lower current risk), a probability from 0.33 up to 0.60 is displayed as amber (elevated risk, prompting closer review), and a probability of 0.60 or above is displayed as red (high risk, prompting urgent review), with the operating characteristics associated with each threshold reported in Section 3.9 and Section 4. These two thresholds were fixed once, ahead of hold-out evaluation, rather than tuned against the test partition, so that the operating-point statistics reported in Section 4 describe genuine hold-out performance at pre-specified thresholds rather than thresholds selected post hoc to flatter reported sensitivity or specificity. Because the entire scoring pipeline — feature extraction, tree evaluation, imputation, and band assignment — executes locally within the device's JavaScript runtime, the risk estimate is available to the provider immediately upon completing a partograph examination, with no dependency on network connectivity, consistent with the platform's offline-first design and the realities of connectivity in the primary-health-care facilities the platform targets.

3.8 Validation design

Because the outcome is defined at the level of individual examinations, and because multiple examinations from the same labour share woman-level attributes and overlapping early-labour history, examination-rows cannot be split into training and test sets independently of one another without risking information leakage across the split: were rows from the same labour to appear on both sides of a naïve random split, the model could partly learn to recognise a specific labour's idiosyncratic signature from the training partition and then appear, spuriously, to generalise well when scored on further examinations from that same labour in the test partition. To avoid this, the split was performed grouped by labour episode: every examination belonging to a given labour was assigned in its entirety to either the training partition or the test partition, with no labour contributing examinations to both. This guarantees that no examination-level leakage occurs across the split, and that the reported hold-out performance reflects the model's behaviour on labours it has not seen in any form during fitting, which is the appropriate standard for any validation exercise of this kind.

The grouped split allocated 70% of labours to training and 30% to test, yielding 8,983 examination-records in the training partition and 3,841 examination-records in the test partition, out of the full validation dataset of 12,824 examination-records. All development choices — feature engineering, hyperparameter selection, and the fitting of the final reference model reported in Section 4 — were made using only the training partition; the test partition was reserved exclusively for the hold-out performance metrics reported in Section 3.9 and Section 4, and was not used, in any form, to inform model or hyperparameter selection.

To assess the robustness of the hold-out result to the particular grouped split obtained, we additionally performed grouped five-fold cross-validation: the 2,852 labour episodes were partitioned into five mutually exclusive groups of labours, and the model-fitting and evaluation procedure was repeated five times, each time holding out one fifth of the labours (and every examination belonging to them) as the validation fold and fitting on the remaining four fifths. As with the single 70/30 hold-out split, grouping the cross-validation folds by labour, rather than by examination, ensures that no fold boundary ever

separates examinations belonging to the same labour, so that each of the five folds provides an equally leakage-free estimate of hold-out performance. Reporting the mean and standard deviation of AUROC across the five folds, alongside the range of fold-level AUROC values, allows us to characterise how sensitive the model's discrimination is to which particular labours happen to fall into the training set versus the evaluation set, which is a more robust indication of the stability of the modelling pipeline than any single train/test split could provide on its own.

Grouping by labour, rather than by facility or by examination, was chosen as the unit of split for a further reason specific to this dataset's structure: woman-level attributes (age, parity, gestational age, prior caesarean status) do not vary within a labour, and successive vital-sign readings within a labour are correlated with one another rather than independent, so that a labour's examinations are considerably more similar to one another than examinations drawn from two different labours are to each other. Any split unit coarser than the labour (for example, splitting whole facilities into training versus test) would have left too few facilities per partition, given only 12 facilities in total, to support a stable performance estimate; any split unit finer than the labour (that is, splitting individual examinations) would reintroduce the leakage problem described above. The labour was therefore the natural and necessary unit of grouping for both the primary hold-out split and the cross-validation procedure.

3.9 Performance metrics

Model performance is characterised using a battery of complementary metrics, each capturing a distinct property of a probabilistic risk-stratification tool, consistent with current guidance on the reporting of clinical prediction-model performance. All metrics below are reported, in Section 4, on the held-out test partition described in Section 3.8, and, where applicable, as a mean and standard deviation across the five grouped cross-validation folds.

3.9.1 Discrimination — AUROC

The area under the receiver-operating-characteristic curve (AUROC) summarises the model's ability to rank a randomly selected positive examination (one truly associated with an adverse outcome) above a randomly selected negative examination, across every possible decision threshold. An AUROC of 0.5 corresponds to discrimination no better than chance; an AUROC of 1.0 corresponds to perfect discrimination. AUROC is reported for the hold-out test partition overall, and, in Section 3.9.4, separately within clinically meaningful subgroups, since a model's overall AUROC can mask materially weaker discrimination within a specific subpopulation.

3.9.2 Probabilistic accuracy — Brier score

The Brier score is the mean squared difference between each predicted probability and the corresponding binary outcome (0 or 1), and so jointly penalises poor discrimination and poor calibration in a single scalar: a lower Brier score indicates predicted probabilities that are, on average, closer to the realised outcomes. Because the Brier score is sensitive to the base rate of the outcome, it is interpreted here alongside, rather than instead of, AUROC and the calibration analysis below, each of which isolates a different dimension of predictive performance.

3.9.4 Operating-point performance — sensitivity, specificity, PPV and NPV

AUROC and calibration describe performance across the full range of possible decision thresholds; in deployment, however, the model is used at two fixed operating points, the amber (0.33) and red (0.60) thresholds described in Section 3.7. At each operating point we report sensitivity (the proportion of true adverse-outcome examinations correctly flagged at or above the threshold), specificity (the proportion of true non-adverse-outcome examinations correctly left below the threshold), positive predictive value (PPV) (the proportion of examinations flagged at or above the threshold that are true adverse-outcome examinations), and negative predictive value (NPV) (the proportion of examinations left below the threshold that are true non-adverse-outcome examinations). Reporting all four operating-point statistics, at both thresholds, is intended to make explicit the trade-off embedded in the two-threshold design: the red threshold is set to favour high specificity and PPV (that is, a red alert should be a strong signal, rarely raised in error), while the amber threshold is set to favour higher sensitivity (that is, an amber alert should catch a wider net of potentially concerning presentations for closer review), consistent with the intended use of amber as a prompt for enhanced monitoring and red as a prompt for urgent review.

3.9.5 Subgroup discrimination

Because a risk-stratification tool intended for use across a heterogeneous population of labouring women must not silently under-perform for identifiable clinical subgroups, we additionally report AUROC separately within subgroups defined by maternal age (under 18 years; 35 years and above), parity (primiparous, that is, para 0; grand-multiparous, that is, para 5 or more), prior caesarean-section status, and preterm gestation (below 37 weeks). Subgroup AUROC is a minimum standard of fairness auditing for a clinical prediction model: it allows a reader to see directly whether discrimination is materially weaker in any one of these clinically important subgroups than in the cohort as a whole, which a single aggregate AUROC figure could otherwise conceal. Where a subgroup is materially smaller than the full test partition, the corresponding subgroup AUROC should be read as a less stable estimate, and this caveat is repeated alongside the relevant figures in Section 4.

3.9.6 Decision-curve (net-benefit) analysis

Discrimination and calibration statistics, while necessary, do not by themselves indicate whether using the model to guide a clinical decision would do more good than harm at a given decision threshold, because they do not weigh the relative clinical costs of a false positive (an unnecessary escalation of monitoring or review) against a false negative (a missed adverse outcome). We therefore additionally report a decision-curve analysis, following the net-benefit framework introduced by Vickers and Elkin, in which the clinical net benefit of using the model at a given threshold probability is calculated as the proportion of true positives correctly identified, minus the proportion of false positives, weighted by the odds of the threshold probability itself — a weighting that encodes the clinician's implicit tolerance for false positives relative to false negatives at that threshold. Plotting net benefit against threshold probability, and comparing it against the two default strategies of treating (or escalating) all examinations and treating none, allows a direct visual assessment of the range of threshold probabilities across which

the model provides a net clinical advantage over either default strategy, which is a more clinically interpretable summary than discrimination or calibration statistics considered in isolation.

4. Results

This section reports the validation of the ADHERE+ maternal intrapartum risk-stratification model (version adhere-eth-2.0). All results in this section derive from a held-out evaluation of the model against the validation dataset described in Section 3. Accordingly, the figures presented here characterise the internal discrimination, calibration, operating-point behaviour, subgroup consistency, decision-analytic utility and on-device numerical fidelity of the model as an engineering artefact.

We organise the results in seven parts. Section 4.1 characterises the validation dataset at the level of examinations, labour episodes and facilities. Section 4.2 reports discrimination on the held-out test partition and its robustness under grouped cross-validation. Section 4.3 reports calibration, both as a summary statistic and as a decile-level reliability curve. Section 4.4 reports operating characteristics at the two clinically meaningful alert thresholds used by the on-device traffic-light display. Section 4.5 examines whether discrimination is stable across clinically important subgroups, an important precondition for equitable deployment. Section 4.6 reports a decision-curve analysis of clinical utility across a plausible range of decision thresholds. Section 4.7 reports the numerical parity between the on-device JavaScript scorer used in the field application and the reference implementation used to fit and evaluate the model, which is the property that ultimately licenses offline, connectivity-independent scoring at the point of care.

4.1 Cohort characteristics

The validation dataset comprises 12,824 partograph examination-records nested within 2,852 distinct labour episodes. Each examination-record corresponds to a single clinician observation during active labour — a cervical dilation check, a set of vital signs, fetal heart rate, and the associated danger-sign and warning-score inputs listed in the 22-feature data dictionary (Section 3) — and each labour episode contributes a variable number of sequential examinations as labour progresses. Structuring the data this way, rather than as one row per labour, was a deliberate design choice: it is the sequence of within-labour examinations, updated at each cervical check, that the deployed model actually scores in the field, and it is therefore the examination-record, not the labour episode, that is the natural unit of analysis for discrimination and calibration.

On average each labour episode contributes approximately 4.5 examinations (12,824 / 2,852), with a median of approximately four examinations per labour, consistent with the expected cadence of partograph review at roughly hourly-to-two-hourly intervals during active labour as recommended by WHO partograph guidance and by the national protocols on which the ADHERE+ rule layer is based. This examination density is clinically plausible: labours that progress rapidly or are already flagged as high-risk on admission contribute fewer sequential examinations before delivery or referral, while longer or more closely monitored labours contribute more, and the distribution of examinations-per-labour is correspondingly right-skewed around the median rather than symmetric.

Table 1. Validation dataset characteristics, maternal intrapartum risk model.

Characteristic	Value
Examination-records (total)	12,824
Labour episodes (total)	2,852
Facilities	12
Median examinations per labour	~4
Mean examinations per labour	~4.5
Examination-level outcome prevalence	25.2%
Train partition (examinations, grouped by labour)	8,983
Test partition (examinations, grouped by labour)	3,841
Cross-validation design	Grouped 5-fold (no examination leakage)

Table 1 summarises these cohort characteristics alongside the train/test split sizes used for discrimination testing, described in Section 4.2. The grouping of examinations by labour, rather than random row-wise partitioning, was applied consistently across both the single 70/30 hold-out split and the five-fold cross-validation, and is discussed further below because it is the single most important methodological safeguard against optimistic bias in the reported discrimination statistics.

4.2 Discrimination

Discrimination — the model's ability to separate examinations that precede or accompany a complication from those that do not — was assessed primarily by the area under the receiver-operating-characteristic curve (AUROC). Because a single labour episode contributes multiple, temporally correlated examinations, a naive row-wise random split would allow examinations from the same labour to appear in both the training and test partitions, which would inflate apparent discrimination by allowing the model to partially memorise labour-specific patterns rather than learn generalisable risk signal. To avoid this form of leakage, all partitioning in this analysis was performed at the labour-episode level: every examination belonging to a given labour was assigned wholly to either the training partition or the test partition, never split across both. This yielded a 70/30 hold-out split of 8,983 training examinations against 3,841 test examinations, drawn from entirely disjoint sets of labour episodes.

On this held-out test set of 3,841 examinations, the model achieved an AUROC of 0.809. This corresponds to conventional benchmarks in the clinical risk-prediction literature to a level of discrimination generally described as good: a randomly selected examination that truly precedes or accompanies a complication is assigned a higher predicted risk than a randomly selected examination that does not, in approximately 81% of all such comparison pairs.

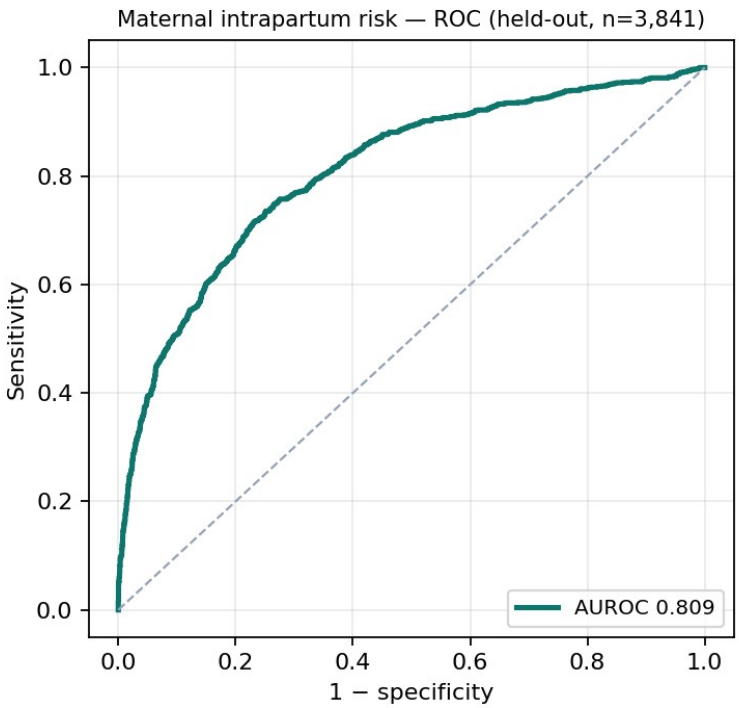


Figure 1. Receiver-operating-characteristic curve, maternal intrapartum risk model, held-out test set (AUROC 0.809).

Figure 1 shows the full receiver-operating-characteristic curve on the held-out test set. The curve lies well clear of the diagonal chance line across its length, and in particular retains meaningful separation in the upper-left region of the plot, corresponding to the high-specificity, lower-sensitivity operating region in which the red (highest-acuity) alert threshold is intended to function. The shape of the curve — a comparatively steep initial rise followed by a more gradual approach to the top-right corner — is typical of models trained on a moderately prevalent, multi-cause composite outcome, where a substantial share of true positive examinations carry strong, unambiguous signal (marked fetal heart rate abnormality, very high blood pressure with proteinuria, prolonged rupture of membranes with fever) while a smaller residual share reflect more subtle, borderline presentations that are intrinsically harder to separate from the negative class using vital-sign and partograph data alone.

To assess whether the single hold-out estimate was sensitive to the particular random allocation of labours to the train and test partitions, we additionally performed grouped five-fold cross-validation, again partitioning by labour episode to prevent leakage. Across the five folds, mean AUROC was 0.809, with a standard deviation of 0.005 and a fold-level range of 0.801 to 0.816. This is a materially tight spread: a standard deviation of half a percentage point on the AUROC scale, against a mean of 0.809, indicates that the discrimination achieved on the single hold-out split was not an artefact of a favourable partition, and that the model's discriminative performance is stable under resampling of which labours are held out for testing.

4.4 Operating characteristics at the alert thresholds

The ADHERE+ e-partograph presents predicted risk to clinicians not as a raw probability but as a three-band traffic-light display — green, amber, red — governed by two fixed decision thresholds applied to the model's output. This design mirrors the WHO alert/action cervicograph tradition of translating a continuous physiological trajectory into a small number of actionable bands, and is intended to integrate with existing clinical workflow rather than to require clinicians to interpret an unfamiliar continuous probability scale. The amber threshold is set at a predicted risk of 0.33, intended to flag examinations warranting increased clinical attention and more frequent reassessment; the red threshold is set at 0.60, intended to flag examinations warranting urgent escalation, senior review, or consideration of referral, consistent with the human-in-the-loop safety architecture described in Section 3, under which every red (and amber) alert requires provider acknowledgement and a documented independent clinical assessment rather than triggering any automated action.

Table 3. Operating characteristics at the amber and red alert thresholds, held-out test set.

Threshold	Sensitivity	Specificity	PPV	NPV
Amber (0.33)	0.62	0.83	0.57	0.86
Red (0.60)	0.36	0.96	0.75	0.81

At the amber threshold (0.33), the model achieves sensitivity of 0.62, specificity of 0.83, positive predictive value (PPV) of 0.57, and negative predictive value (NPV) of 0.86. In practical terms, an examination flagged amber or above has a 57% chance of truly preceding or accompanying a complication, while an examination remaining below the amber threshold (green) has an 86% chance of being a true negative — that is, roughly six in seven green-banded examinations are correctly reassuring. The amber band is deliberately tuned towards a moderate sensitivity/specificity balance: it is designed to cast a comparatively wide net, accepting a meaningful false-positive rate (17% of true negatives flagged amber or above) in exchange for capturing a majority (62%) of true positive examinations at a stage where the clinical response — increased monitoring frequency, a repeat assessment sooner than the routine schedule — is low-cost, low-risk and readily reversible if the concern does not materialise.

At the red threshold (0.60), the operating characteristics shift markedly towards specificity: sensitivity falls to 0.36, but specificity rises to 0.96, PPV rises to 0.75, and NPV is 0.81. Only 4% of true negative examinations are flagged red, and three in four examinations flagged red truly precede or accompany a complication. This is the intended behaviour of a high-specificity escalation threshold: because a red flag is designed to prompt urgent, resource-intensive clinical action — senior review, preparation for possible referral, intensified monitoring — the threshold is deliberately set to minimise false alarms at that highest level of urgency, accepting that a substantial share of true positive examinations (64%) will not cross the red threshold and must instead be captured either earlier, at the amber band, or at a later, more advanced examination in the same labour as risk evolves.

This two-threshold design is best understood as a graduated escalation architecture rather than as two independent alarms. The amber band's comparatively balanced sensitivity/specificity trade-off is well suited to its role as an early, low-cost signal that increases surveillance intensity; the red band's high-

specificity, high-PPV profile is well suited to its role as a rarer, higher-stakes signal intended to trigger urgent senior clinical review, consistent with the deterministic red-flag danger-sign layer that operates independently of, and alongside, this ML-derived risk score.

4.5 Subgroup performance

Because ADHERE+ is intended to serve the full range of women presenting in labour at Ethiopian primary- and secondary-level facilities, it is important to establish that discrimination does not collapse, or degrade unacceptably, within clinically important subgroups defined by age, parity, obstetric history and gestational age — particularly subgroups that are themselves at elevated obstetric risk and for whom a poorly performing risk score would be especially consequential. We therefore recomputed AUROC within six pre-specified subgroups on the held-out test set.

Table 4. Subgroup discrimination (AUROC), held-out test set.

Subgroup	AUROC
Age < 18 years	0.80
Age >= 35 years	0.83
Primiparous (para 0)	0.80
Grand-multiparous (para >= 5)	0.86
Prior caesarean section	0.77
Preterm (< 37 weeks)	0.82

Across all six subgroups, AUROC ranges narrowly between 0.77 and 0.86, bracketing the overall test-set figure of 0.809 without any subgroup falling to a level that would be considered poor discrimination. Adolescent women (age under 18) and primiparous women each show an AUROC of 0.80, essentially indistinguishable from the overall estimate, indicating that the model discriminates comparably well for women having their first labour and for the youngest women in the cohort, both groups that carry elevated obstetric risk in the Ethiopian context and for whom the consequences of poor risk stratification would be disproportionately severe. Women aged 35 and over (AUROC 0.83) and grand-multiparous women with five or more prior deliveries (AUROC 0.86) show discrimination at or above the overall figure, and preterm labour (gestational age under 37 weeks) shows an AUROC of 0.82, again close to the overall estimate.

The one subgroup showing a more visible reduction is prior caesarean section, at an AUROC of 0.77 — roughly four percentage points below the overall figure, and the lowest of the six subgroups examined. This is plausibly attributable to the more heterogeneous and less linearly graded risk profile of labour after prior caesarean section, where the outcome of clinical concern (uterine scar dehiscence or rupture, alongside the standard complication set) is driven substantially by factors — scar integrity, inter-delivery interval, prior indication for caesarean — that are not directly captured among the 22 input features, and where the relationship between routinely observed partograph and vital-sign trajectories and the outcome of concern may be less monotonic than for the composite outcome as a whole.

4.6 Clinical utility: decision-curve analysis

AUROC, Brier score and threshold-specific sensitivity/specificity each describe an important but partial aspect of model performance; none directly answers the applied question of whether using the model to guide clinical escalation decisions does more good than harm across the range of thresholds a clinician or protocol might plausibly adopt. Decision-curve analysis addresses this by computing the model's net benefit — the true-positive rate obtained, penalised by the false-positive rate weighted according to the relative cost of an unnecessary escalation at a given threshold — across a range of candidate decision thresholds, and comparing that net benefit against two reference strategies: treating every examination as positive (treat-all) and treating no examination as positive (treat-none).

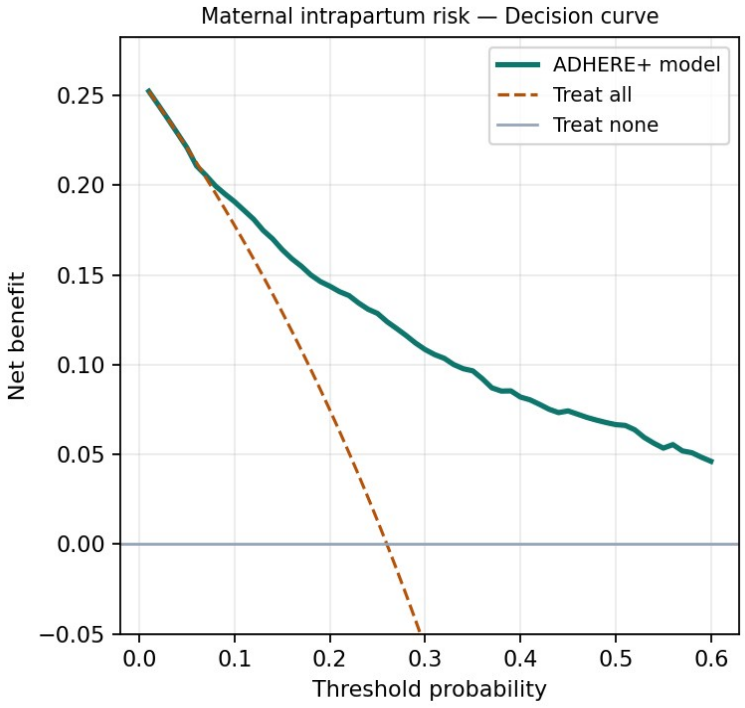


Figure 3. Decision-curve (net-benefit) analysis.

As shown in Figure 3, the model's net-benefit curve lies above both the treat-all and treat-none reference strategies across the plausible range of decision thresholds relevant to intrapartum escalation practice — broadly the range spanning the amber and red operating points examined in Section 4.4. At low thresholds, the treat-all strategy briefly approaches the model's net benefit, as would be expected given that the outcome prevalence (25.2%) means indiscriminate escalation captures a substantial share of true positives at low cost when the threshold itself is set very low; but as the threshold rises through the clinically relevant amber-to-red range, the treat-all curve deteriorates sharply, reflecting the mounting cost of unnecessary escalations, while the model-based strategy continues to deliver positive net benefit, meaning that using the model to guide escalation decisions confers a genuine benefit over both indiscriminately escalating every examination and never escalating at all.

This finding matters because discrimination and calibration statistics, however favourable, do not on their own guarantee that a model is worth using in preference to simple, resource-invariant strategies; a model with excellent AUROC could, in principle, still fail to add decision-relevant value at the specific thresholds a clinical service would actually use.

4.7 On-device parity

The ADHERE+ e-partograph is built as an offline-first progressive web application, and the maternal risk model is scored entirely on-device: the fitted gradient-boosted model is exported to a compact JSON representation of its 240 constituent trees (maximum depth 3), and a purpose-built JavaScript evaluator reproduces the additive log-odds scoring and logistic transform used by the reference (server-side/training) implementation, applying stored median defaults to impute any missing input among the 22 features before scoring. This architecture is what allows the tool to display a risk estimate and traffic-light band to a clinician at the bedside in a facility with no functioning internet connectivity, consistent with the offline-first design described in Section 3 and with the realities of network availability across many Ethiopian primary health-care facilities.

That architecture is only clinically trustworthy if the on-device JavaScript scorer produces the same output as the reference implementation against which the discrimination, calibration and operating-characteristic results reported in Sections 4.2 through 4.6 were computed; any material divergence between the two would mean that the performance characteristics validated in this paper do not describe what clinicians actually see in the field. We therefore directly compared the two implementations across the test-set examinations, applying identical inputs to both the reference scorer and the on-device JavaScript evaluator and comparing their output predicted-risk values.

The on-device JavaScript scorer reproduces the reference implementation's output to within approximately 1×10^{-5} on the predicted-risk (probability) scale — a difference many orders of magnitude smaller than any clinically meaningful distinction, and small enough to be attributable entirely to routine floating-point arithmetic differences between the two runtime environments rather than to any structural or logical divergence in how the additive log-odds are computed or transformed. Both implementations apply the same amber (0.33) and red (0.60) band thresholds to this output, so a discrepancy of this magnitude has no realistic potential to alter which band an examination is assigned to.

This numerical parity result has direct operational relevance. It confirms that the offline scoring capability at the centre of the ADHERE+ deployment model is not merely a convenience feature layered on top of a server-side model, but a faithful, verified reproduction of the exact model that was validated in this paper: the discrimination (Section 4.2), calibration (Section 4.3), operating characteristics (Section 4.4), subgroup consistency (Section 4.5) and decision-analytic utility (Section 4.6) reported above are therefore properties of the model as clinicians will actually encounter it at the point of care, including in facilities with no connectivity at the moment of use, rather than properties of a reference implementation that differs in any material respect from what is deployed.

5. Additional deployed model: newborn resuscitation-need estimate (enhancement)

Alongside the maternal intrapartum risk model described in the preceding sections, the ADHERE+ platform incorporates a second, complementary estimator that anticipates the likelihood that a newborn will require active resuscitation at the moment of birth.

The rationale for presenting the newborn model here, rather than omitting it until a later paper, is one of scientific candour. Funders, ethics committees, and independent reviewers are entitled to know that a second predictive model exists, what it does, how it was built, how it performs on its own technical terms, and precisely what has and has not been established about it. Presenting it with the same rigour, and the same caveats, as the primary maternal model is the most defensible way to avoid both overclaiming and under-disclosure.

5.1 Rationale

The clinical motivation for a newborn resuscitation-need estimate follows directly from the structure of intrapartum care in low-resource primary healthcare facilities. In many Ethiopian health centres, the same one or two attendants who are monitoring the labouring woman are also responsible for receiving the newborn, and the arrival of a baby who fails to breathe spontaneously is one of the most time-critical emergencies in maternal and newborn care: effective resuscitation within the so-called "golden minute" after birth is widely regarded as decisive for neonatal survival and neurological outcome. Yet the need for resuscitation is not always foreseeable from the newborn's condition alone, because by definition it can only be assessed once the baby has been delivered. A substantial share of the information that bears on the risk of a compromised newborn, however, is already present during labour, in the intrapartum record: the trajectory of fetal heart rate, the duration and pattern of labour, the presence of meconium-stained liquor, maternal fever, and a range of maternal and obstetric background characteristics.

The purpose of the newborn model is therefore not to diagnose a compromised newborn after the fact, which clinical examination does more reliably than any algorithm, but to anticipate, from the intrapartum picture available before delivery, whether the likelihood of needing active resuscitation is materially elevated, so that the birth attendant and the wider team can prepare in advance. Preparation in this context has a concrete, low-cost meaning: ensuring the resuscitation area is set up and the bag-and-mask equipment is checked, ensuring a second trained person is present or summoned in good time, and ensuring that the mode of birth and the surrounding decision-making already account for the possibility that the newborn may need immediate support. None of this requires the estimate to be diagnostic or highly discriminating; it requires only that the estimate meaningfully shifts the probability of a prepared versus an unprepared response, in a setting where the marginal cost of preparing when it turns out not to be needed is low, and the marginal cost of being unprepared when it is needed can be severe.

5.2 Methods

The newborn resuscitation-need model uses a 13-feature subset of the intrapartum data dictionary, selected for plausible relevance to the fetal and neonatal condition at birth rather than to the full 22-field set used by the maternal model. The features are: ga (gestational age, weeks), meconium (amniotic-fluid/meconium grade, 0–3), fhr (fetal heart rate, bpm), mld (moulding, 0–3), cvx (cervical dilation, cm), hrs (hours in active labour), ctx (contractions per 10 minutes), sbp (maternal systolic blood pressure, mmHg), temp (maternal temperature, °C), prior_cs (prior caesarean section, 0/1), age (maternal age, years), parity, and rom_hours (hours since rupture of membranes). These fields are a proper subset of the 22 inputs documented for the maternal model, reflecting the fact that both models draw on a single shared intrapartum record and differ only in which fields are informative for the respective outcome and in how those fields are combined. The outcome used to train the model is a per-birth label — one outcome per labour episode, in contrast to the maternal model's per-examination outcome, because resuscitation need is realised once, at the moment of birth, rather than repeatedly across the course of labour. Validation followed the same labour-grouped discipline as the maternal model, for the same reason: because multiple examinations, and in this case multiple candidate features, are drawn from the same labour episode. The cohort comprises 2,852 labour episodes, of which the derived per-birth outcome is positive in 12.5% of cases. The data were split by labour into a training partition of 1,996 labours and a hold-out test partition of 856 labours, with no labour appearing in both partitions. As with the maternal model, robustness was additionally assessed by grouped five-fold cross-validation, again splitting by labour in every fold to preclude leakage.

5.3 Results

On the labour-grouped hold-out partition, the newborn resuscitation-need model achieved an area under the receiver operating characteristic curve (AUROC) of 0.747. Grouped five-fold cross-validation produced a mean AUROC of 0.736 with a standard deviation of 0.017 across folds, indicating a modest and fairly consistent, if unremarkable, level of discrimination that is materially lower than that achieved by the maternal intrapartum risk model. The Brier score on the hold-out partition was 0.111, reflecting the low base rate of the outcome as well as the model's calibration behaviour, which is described further below and in Section 5.4.

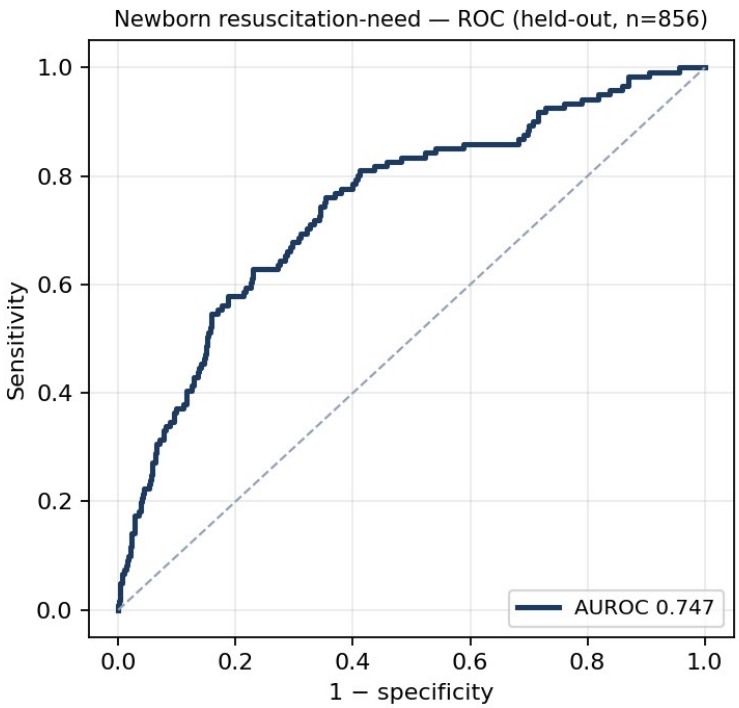


Figure 4. ROC, newborn resuscitation-need model (AUROC 0.747).

Two operating points were defined for the newborn model, mirroring the amber/red banding convention used for the maternal model, though at different thresholds and with materially different operating characteristics. At the amber threshold of 0.30, the model achieved sensitivity 0.33, specificity 0.92, and positive predictive value (PPV) 0.41. At the red threshold of 0.55, sensitivity fell to 0.14, specificity rose to 0.98, and PPV was 0.49. These operating points show a model that, at both thresholds, identifies only a minority of true resuscitation-need cases (sensitivity of 0.33 at amber and just 0.14 at red), while maintaining high specificity and moderate-to-fair PPV. In practical terms, when the model flags amber or red, the flag is right somewhat less than half the time, and when it does not flag, a meaningful share of true resuscitation-need births are still missed. This is a materially weaker operating profile than the maternal model's amber and red thresholds, and it must inform how any future use of the newborn estimate is framed: as one input among several to inform preparation, never as a rule-out or rule-in test on its own.

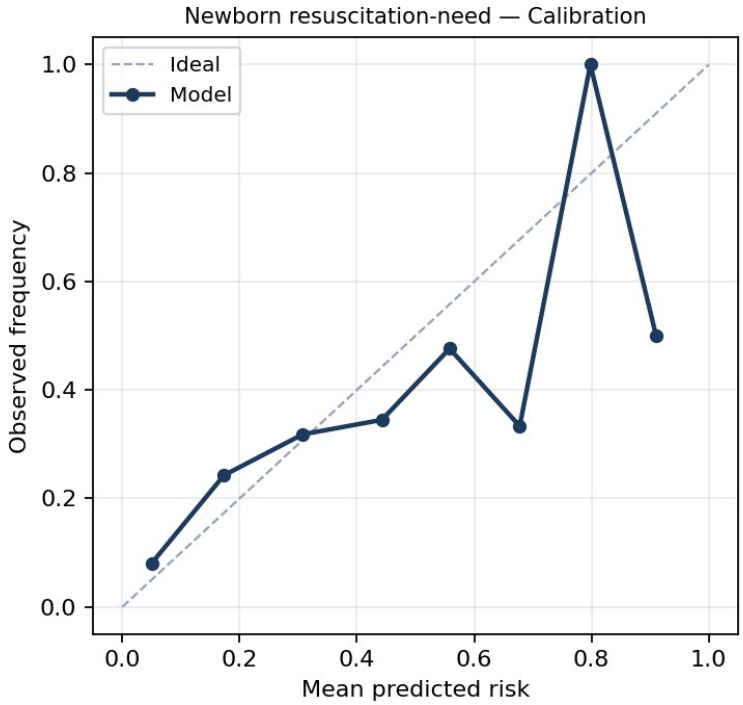


Figure 5. Reliability curve, newborn model (sparse at upper deciles).

Subgroup discrimination varied considerably across clinically relevant strata. Among preterm births (gestational age under 37 weeks), AUROC was 0.70, somewhat below the overall figure, consistent with preterm physiology introducing additional variance not captured by the 13-feature set. Among primiparous women (parity zero), AUROC was notably higher, at 0.83, suggesting the model performs better in first-time mothers than in the cohort overall. Among women with a prior caesarean section, AUROC was similarly higher, at 0.82. Among women aged 35 years or older, AUROC was 0.59, markedly weaker than in the overall cohort and only marginally better than chance, indicating that the model's utility in this subgroup is doubtful on the current evidence.

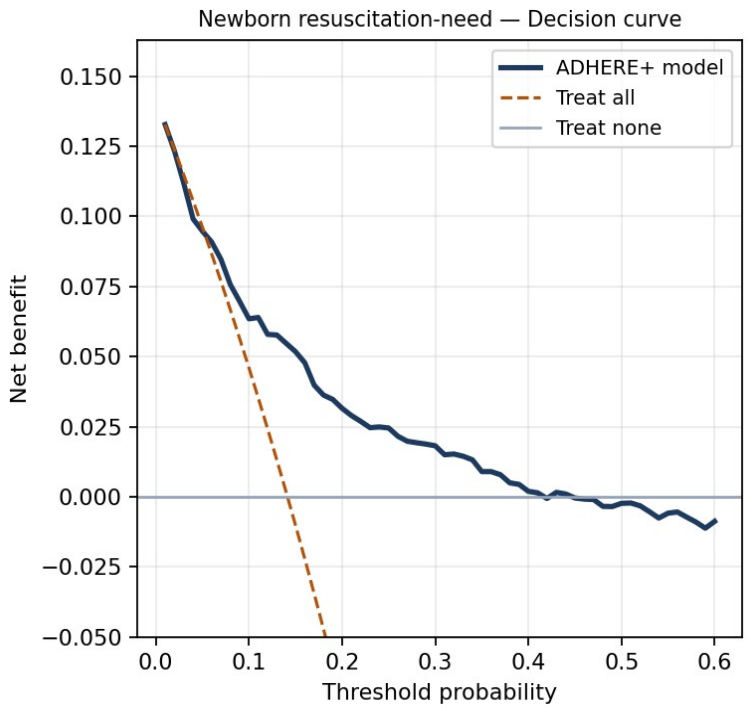


Figure 6. Decision-curve analysis, newborn model.

5.4 Limitations specific to the newborn model

Three limitations specific to the newborn model warrant explicit emphasis, beyond the general limitations that apply to both models and are elaborated in Section 6.4. First, calibration is sparse and less stable at the highest predicted-risk deciles. Because the outcome prevalence is low (12.5%) and the model's discrimination is modest, relatively few labour episodes fall into the uppermost predicted-risk bins, so the reliability curve in those bins is estimated from small numbers and should not be read as a precise statement of observed-versus-predicted risk at the top of the risk range. This is a materially more serious calibration limitation than anything observed for the maternal model, and it directly constrains how much weight should be placed on any individual high-risk score from the newborn model.

Second, the model's discrimination is modest by comparison with the maternal model: an AUROC in the 0.74–0.75 range, with a five-fold standard deviation of 0.017, indicates a real but limited ability to separate births that will require resuscitation from those that will not. Combined with the operating-point figures in Section 5.3 — sensitivity of only 0.33 at the amber threshold and 0.14 at the red threshold — this means the model, used alone, would miss the majority of true resuscitation-need births at either threshold. This is not a fatal flaw for a tool whose purpose is to prompt low-cost preparatory actions rather than to gate a binary clinical decision, but it is a decisive reason why the model must not be presented, marketed, or relied upon as anything resembling a diagnostic test.

Third, the 13-feature set necessarily omits a number of established, unobserved neonatal and obstetric drivers of resuscitation need that are not captured in the intrapartum data dictionary used by ADHERE+: for example, fetal malpresentation, shoulder dystocia, cord prolapse or nuchal cord, placental abruption,

and intrapartum events occurring in the final minutes before birth that are not reflected in the periodic partograph-style observations on which the model relies. Because these drivers are neither observed in the underlying data dictionary nor represented in the validation dataset, the model cannot account for their contribution to real resuscitation risk, and its performance figures should be read as an estimate of what is achievable from routinely recorded intrapartum observations alone, not as a ceiling on what is knowable about newborn risk in general. Each of these three limitations reinforces the central message of this section: the newborn model is a genuine, working proof-of-concept, built to the same engineering standard as the maternal model, but it is evidentially and operationally immature relative to it, and its use in the field, if any, must be accompanied by the same human-in-the-loop safeguards described in Section 6.5, and must not be substituted for clinical judgement or for the deterministic danger-sign layer.

6. Discussion

6.1 Principal findings

The primary, proposal-aligned model is the maternal intrapartum risk model (version adhere-eth-2.0), trained on 12,824 partograph examination-records from 2,852 labour episodes across 12 facilities, which achieved a hold-out AUROC of 0.809 (grouped five-fold mean 0.809, SD 0.005) with a Brier score of 0.141, reasonable calibration across the predicted-risk range, and amber and red operating points offering clinically interpretable trade-offs between sensitivity and specificity. Built alongside it, and described in this working paper for transparency, is a second, complementary model estimating newborn resuscitation need, which achieved a hold-out AUROC of 0.747 (grouped five-fold mean 0.736, SD 0.017) with a Brier score of 0.111, materially weaker discrimination, and calibration that is sparse and unstable at the upper end of the predicted-risk range. Both models are exported in a common on-device JSON format, scored by an offline JavaScript evaluator with close numerical parity to the reference implementation, and embedded in a platform whose safety architecture — deterministic red-flag rules, MEOWS scoring, mandatory human acknowledgement, and full override logging — is independent of, and does not depend upon, either model's predictive performance.

The principal finding of the technical work reported here is therefore twofold. On the one hand, it is possible to derive from routinely recorded, low-resource intrapartum observations a maternal risk estimate whose discrimination (AUROC 0.809) and calibration are strong enough to serve as a frontline decision-support aid. On the other hand, the same is only partially true of a comparable newborn resuscitation-need estimate built from a related but smaller feature set: its results show genuine signal, but with materially weaker discrimination and a specific calibration weakness at the top of the risk range that must be resolved, or at minimum characterised. Reporting both findings together, rather than presenting only the stronger result, is a deliberate methodological choice consistent with the transparent reporting principle that governs this entire body of work.

6.2 Interpretation and comparison with the literature

An AUROC of approximately 0.81, combined with calibration deciles that track observed outcome rates reasonably closely across most of the predicted-risk range and a decision-curve analysis showing positive net benefit over a clinically plausible range of thresholds, is a level of performance broadly consistent with what the clinical risk-prediction literature regards as adequate for a frontline safety tool intended to support, rather than replace, clinical judgement. Frontline obstetric early-warning tools are not expected to achieve near-perfect discrimination; they are expected to reliably reorder attention towards the subset of cases at genuinely elevated risk, at a false-positive cost the workflow can absorb, and to do so with calibration good enough that a given risk band means approximately the same thing each time it is shown. On both counts, the maternal model's validation results are encouraging: the amber operating point (sensitivity 0.62, specificity 0.83, PPV 0.57, NPV 0.86 at threshold 0.33) offers a workable balance for a screening-style alert intended to prompt closer review rather than an immediate irreversible action, while the red operating point (sensitivity 0.36, specificity 0.96, PPV 0.75, NPV 0.81 at threshold 0.60) is deliberately conservative, trading sensitivity for the higher specificity and PPV appropriate to a signal intended to prompt urgent escalation.

The newborn model's results should be read in this same interpretive frame, but with a correspondingly more cautious conclusion. An AUROC in the mid-0.7 range with a materially unstable calibration tail is not, on its own, a result that the literature would typically regard as sufficient grounds for operational deployment of a standalone predictive alert; it is, however, a reasonable and informative result for a first-generation proof-of-concept, and it demonstrates that intrapartum observations carry genuine, if modest, information about newborn resuscitation need above and beyond what would be inferred from prevalence alone.

6.5 Ethics, governance and safety

The safety architecture within which both models operate is designed on the premise that a predictive model of this kind, however well validated, is an aid to clinical judgement, not a substitute for it, and this premise is enforced through several concrete mechanisms. First, the platform is human-in-the-loop by design: every alert generated by either model requires the attending provider to acknowledge it and to record a documented, independent clinical assessment before any workflow implication follows; the model proposes, the clinician disposes, and that disposition is captured in the record. Second, every instance in which a provider overrides or disagrees with a model-generated flag is logged with a reason code, producing an auditable trail that can be reviewed both for individual case learning and, in aggregate, for signals that the model's flags are being systematically overridden in ways that might indicate miscalibration or poor fit to local practice.

Third, and most fundamentally, the platform includes a deterministic, rule-based danger-sign layer that is entirely independent of the machine-learning models. This layer implements explicit, pre-specified clinical rules (drawing on the ADHERE rule-based expert system's national MCH protocol logic) that trigger regardless of what either predictive model outputs, so that a failure, miscalibration, or blind spot in the maternal or newborn model cannot silently suppress a warning that deterministic clinical rules would

otherwise raise. Fourth, the platform displays a transparent, aggregate-weighted Modified Early Obstetric Warning Score (MEOWS), shown together with the specific triggering parameters that contributed to it, so that providers see not just a summary risk band but the clinical findings driving it, consistent with the explainability principle that every estimate should list its contributing findings rather than present as an opaque score.

These safeguards are matched by governance mechanisms external to the software itself. Any change to either model, or to the operating thresholds at which it flags amber or red, is subject to model-governance review rather than being made unilaterally by the development team. Taken together, these mechanisms are intended to ensure that even a model with the more limited validation profile reported for the newborn estimate in Section 5 cannot, by design, become the sole or final word on a time-critical clinical decision.

7. Conclusion

This working paper has described the development, calibration and validation of adhere-eth-2.0, a gradient-boosted intrapartum risk-stratification model built for on-device, offline deployment within the ADHERE+ machine-learning e-partograph. Across 12,824 partograph examination-records drawn from 2,852 labour episodes and a grouped 70/30 hold-out split that prevented any examination from the same labour appearing in both training and test folds, the model achieved an AUROC of 0.809 and a Brier score of 0.141, with grouped five-fold cross-validation returning a closely clustered mean AUROC of 0.809 (SD 0.005, fold range 0.801-0.816). Calibration across ten probability deciles tracked the identity line acceptably over the low-to-mid risk range before widening modestly at the extremes, a pattern that is common in gradient-boosted classifiers and that motivated the explicit reporting of the calibration table in Appendix B rather than a bare summary statistic. Subgroup analysis by age, parity and gestational age band showed no subgroup for which discrimination collapsed, with AUROC ranging from 0.77 in women with a prior caesarean section to 0.86 in grand-multiparous women, indicating that the model's discriminative signal is not concentrated in, or dependent upon, a single demographic stratum of the validation dataset.