

Inferring Optimised Pregnancy Conception Dates from Multiple Gestational Age Estimates: A Hybrid Anomaly Detection and Linear Mixed Model Approach

Mercy Chepkirui

`mercy.chepkirui@lstmed.ac.uk`

KEMRI - Center for Global Health Research

Stephanie Dellicour

Liverpool School of Tropical Medicine

Benard Omondi

KEMRI - Center for Global Health Research

Kennedy Maube

KEMRI - Center for Global Health Research

Gerald Ongayo

KEMRI - Center for Global Health Research

Michael Alaw

KEMRI - Center for Global Health Research

Benard Asuke

KEMRI - Center for Global Health Research

Titus Ochieng

KEMRI - Center for Global Health Research

Simon Kariuki

KEMRI - Center for Global Health Research

Martina Oneko

KEMRI - Center for Global Health Research

Hellen Barsosio

KEMRI - Center for Global Health Research

Feiko ter Kuile

Liverpool School of Tropical Medicine

Miriam Taegtmeier

Liverpool School of Tropical Medicine

Maia Lesosky

Research Article

Keywords: Pregnancy dating, GA estimation, anomaly detection, linear mixed model, real world data, machine learning, Isolation Forest, pharmacovigilance, perinatal research

Posted Date: July 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10042682/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Inferring Optimised Pregnancy Conception Dates from Multiple Gestational Age Estimates: A Hybrid Anomaly Detection and Linear Mixed Model Approach

Mercy Chepkirui^{1,2*}, Stephanie Dellicour², Benard Omondi^{1,2}, Kennedy Maube¹, Gerald Ongayo^{1,2}, Michael Alaw¹, Benard Asuke¹, Titus Ochieng¹, Simon Kariuki¹, Martina Oneko¹, Hellen Barsosio^{1,2}, Feiko ter Kuile^{1,2}, Miriam Taegtmeyer², Maia Lesosky³

1. Kenya Medical Research Institute, Centre for Global Health Research, Kisumu, Kenya
2. Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom
3. National Heart and Lung Institute, Imperial College, London, United Kingdom

*Corresponding author: [mercy.chepkirui@lstmed.ac.uk]

Abstract

Background

Accurate gestational age (GA) estimation is fundamental for perinatal research and pharmacovigilance but remains challenging in low-resource settings where routine pregnancy records are inconsistent, early ultrasound is limited, and multiple GA estimates often conflict. This challenge is most acute in large-scale observational studies and pharmacovigilance registries, where the data quality controls available in clinical trials are rarely feasible.

Methods

This study implemented a two-stage approach to derive more reliable pregnancy conception dates using routine pregnancy data from 42 healthcare facilities in Western Kenya. We first applied Isolation Forest to identify and exclude implausible GA estimates, then used a linear mixed-effects model to synthesise the remaining measurements into a single calibrated “best” estimate, accounting for systematic biases and pregnancy-level variation. The model pools information across all pregnancies to generate a calibrated conception date for each individual woman, even when her own measurements are limited.

Results

Integrating multi-modal inputs in each pregnancy synthesised a unified GA estimate, providing a robust dating consensus even in the absence of early ultrasound and variation across dating methods. Anomaly detection revealed data quality was unevenly distributed, with anomaly rates higher among pregnancy complications: 0.5% in term live births versus 5.1% in miscarriages. These discrepancies were primarily driven by data entry errors in antenatal clinic records (46.8%) and last menstrual period (LMP) date recall. The mixed effects model identified varying precision among dating methods. Using early ultrasound as the gold standard (mean gestation: 278.6 days), later ultrasound scans showed progressive bias, increasing from 3.15 days in the second trimester to 6.41 days in the third trimester, but retained high precision (SD: 7.5–8.9 days). Despite this drift, ultrasound markers remained more precise than other methods, such as LMP, Ballard score, foot length and fundal height, which had wider uncertainty spreads (>35 days).

Conclusion

Routine care datasets have the potential to yield reliable GA estimates in the absence of widespread early ultrasound through rigorous data cleaning and statistical modelling. However, translating this approach into real-time clinical practice will require embedding these algorithms into interoperable electronic medical records.

Keywords

Pregnancy dating, GA estimation, anomaly detection, linear mixed model, real world data, machine learning, Isolation Forest, pharmacovigilance, perinatal research

1 Background

Accurate determination of gestational age (GA), established through reliable pregnancy dating, is fundamental to clinical care during pregnancy and essential for perinatal health research. GA informs the timing of prenatal care, guides clinical decision-making regarding interventions, and determines expected delivery date (1). Precise pregnancy dating is critical when evaluating medication safety during pregnancy, as accurate temporal alignment between drug exposures and specific windows of foetal development is required to assess teratogenic risk (2). Reliable GA is needed for outcome ascertainment such as preterm birth, small for gestational age and differentiating between miscarriage and stillbirth. In longitudinal studies, accurate record linkage to pregnancy episodes requires reliable start and the end dates for each pregnancy. Without reliable conception dates, both clinical decision-making and research outcomes may be compromised, potentially affecting maternal and neonatal health (3,4).

Multiple approaches exist for estimating GA, each with distinct accuracy profiles and limitations. First-trimester ultrasound measurement of crown-rump length (performed before 14 weeks gestation) is considered the reference standard when available, with an expected precision of ± 5 -7 days according to international guidelines (1,5). However, accuracy deteriorates substantially when ultrasound is performed in the second trimester (± 10 -14 days) or third trimester (± 21 -30 days) (5). Maternal report of last menstrual period (LMP), while readily available, is subject to recall bias, assumes regular 28-day menstrual cycles with ovulation on day 14, and can yield errors exceeding four weeks in populations with irregular cycles or poor recall (6). Clinical assessment of fundal height demonstrates high inter-observer variability and is substantially affected by maternal body physique, with reported accuracy ranging from ± 14 to ± 28 days (7). Postnatal assessment methods, such as the Ballard score, are used to estimate a newborn's gestational age (between 20 and 44 weeks) by evaluating six neuromuscular and six physical maturity characteristics (8). While particularly useful for assessing premature infants, these postnatal methods generally have an accuracy of ± 14 -21 days and may not correlate well with established prenatal dating methods (9). Collectively, these methodologies produce individual measurement error margins ranging from 5 days to 4 weeks, with systematic biases that may vary by population, clinical setting, and operator skill.

These challenges are particularly pronounced in low- and middle-income countries (LMICs) where systematic barriers compound individual measurement limitations. In Kenya, only 29% of women attend their first antenatal care (ANC) visit during the first trimester, with the median time to first contact approximately four months into pregnancy (10,11). This late presentation constrains opportunities for early, accurate ultrasound dating. Furthermore, ultrasound availability remains concentrated in urban referral health facilities, with most rural primary care

facilities lacking both functional ultrasound equipment and trained operators (12,13). Clinical assessment skills, particularly fundal height measurement, depend critically on quality of training and supportive supervision: resources that multiple evaluations have found to be inadequate and inconsistent across Kenya's chronically under-resourced and decentralised health system (14,15). Healthcare workers in rural facilities often receive limited initial training in obstetric assessment and infrequent opportunities for skill reinforcement or quality assurance. These structural deficits result in pregnancy records characterised by substantial dating uncertainties, late and infrequent measurements, and heavy reliance on less accurate methods such as LMP, fundal height, and second or third trimester ultrasound.

Traditionally, clinicians and researchers resolve conflicting or inconsistent conception dates by applying predefined clinical rules or relying on expert review to select the most plausible estimate (1). While effective for individual cases for clinicians and smaller datasets for researchers, these manual or deterministic approaches do not scale well for research in large pregnancy registries or electronic health records and may inadequately address discrepancies arising from multiple, sometimes contradictory data sources (16).

Novel computational techniques offer promising approaches for handling multiple uncertain measurements in medical datasets. Anomaly detection algorithms, such as the Isolation Forest (IF) method, have demonstrated strong performance in identifying implausible or erroneous observations (17). These algorithms identify anomalies based on the principle that outliers are easier to isolate in feature space than normal observations. They have been applied in diverse healthcare applications including fraud detection in medical insurance claims (18), quality control in medical imaging, and patient monitoring systems for detecting abnormal physiological signals that enable early intervention (19). This approach is well-suited for pregnancy dating anomaly detection because: (1) "ground truth" for true conception dates is rarely available, precluding supervised learning approaches; (2) data quality issues in routine datasets vary substantially due to differences in clinical protocols, data collection systems, and healthcare populations (20,21) requiring flexible data-driven approaches rather than rigid rule-based systems.

While anomaly detection identifies implausible values, it cannot generate an estimated best conception date based on the available information. Once implausible observations are removed, the fundamental challenge remains: how should multiple conflicting but plausible measurements be optimally combined to estimate the underlying best conception date? Resolving this requires a systematic modelling approach to integrate these varying estimates into a single, optimised value for each pregnancy. Ideally, this integration must account for method-specific biases and inherent pregnancy-level variability to yield a calibrated 'best' GA aligned with a rigorous standard, such as first-trimester ultrasound. Moreover, by pooling information across all pregnancies, the model improves individual-level estimates even for women with sparse measurements, making it particularly valuable for pharmacovigilance analyses requiring accurate individual pregnancy timing.

Unlike clinical trials, large-scale pharmacovigilance registries and routine care datasets rarely afford the data quality controls needed for reliable GA estimation. To provide more reliable conception dates for perinatal research and the linkage of pregnancy episode records, this study applied a two-stage methodology designed to address data quality concerns, pregnancy-level variability, and method-specific biases. This approach was implemented using data collected from antenatal and maternity clinics within 42 routine healthcare facilities in Homa Bay County, western Kenya. First, machine learning anomaly detection (IF) was used to identify and exclude implausible GA estimates, thereby improving data quality. Second, a linear mixed effects model was fitted to estimate systematic biases in different GA dating methods relative to first-trimester ultrasound and to derive, for each pregnancy, a model-based conception date integrating information across all available dating methods.

2 Methods

2.1 Study Design and Data Sources

This study was embedded within the MiMBa Pregnancy Exposure Registry (22), implemented across 42 healthcare facilities in Homa Bay County, Kenya. The registry captures longitudinal pregnancy data, including medication exposures, for women aged 15-49 years using digital records collected from the community, outpatient and inpatient departments, antenatal care, and maternity services. The registry was integrated within the Ministry of Health infrastructure, where routine paper-based registers completed by community health promoters and healthcare providers were digitised using ScanForm, an optical character recognition technology (23). Trained data clerks scanned these registers to generate structured electronic records. Data collection spanned from January 2021 to June 2025.

2.1.1 Study Population

Pregnancies were included in the analysis if they met the criteria shown in Figure 1. Each pregnancy record included at least three GA estimates, estimated based on methods such as ultrasound examination, maternal recall of LMP, GA assessments at subsequent antenatal visits, fundal height, foot length, or Ballard score measurement at birth.

We excluded pregnancies with fewer than three GA estimates, as insufficient data points limit reliable conception date estimation. Estimates derived solely from birth outcomes (e.g. Term live births assumed to be 40 weeks) and GA recorded in the maternity register were excluded due to known inconsistencies and inaccuracy. Pregnancies with estimated GA falling outside biologically plausible ranges relative to their reported outcomes were excluded to maintain data quality. Additionally, pregnancies exhibiting discordant outcomes, such as twins where one foetus was recorded as live birth and the other as stillbirth, were excluded due to the complexity of reliable dating discordant twin outcomes. Finally, pregnancies resulting in outcomes such as induced abortion, unspecified pregnancy loss, or those with missing outcome

information were excluded because they lack a defined delivery window, making it difficult to determine the timing of the pregnancy outcome.

2.2 Estimation of GA and Conception Dates

First, the IF algorithm was used for anomaly detection to identify and exclude implausible GA measurements, thereby reducing the impact of random errors and data quality issues on subsequent modelling. In the second stage, systematic bias across gestational age methods was addressed using a linear mixed-effects model. This model (a) estimated method-specific systematic biases by including gestational age method and pregnancy outcome as fixed effects and a pregnancy-level random intercept to account for within-pregnancy correlation; and (b) used the resulting calibrated predictions to derive a single pregnancy-specific “best” GA at the observed outcome, obtained as the empirical best linear unbiased prediction (BLUP) for each pregnancy. These model-based GA estimates were then used to derive conception dates and to quantify method-specific bias and agreement relative to the first-trimester ultrasound measurement.

2.2.1 Stage 1: Anomaly Detection Using Isolation Forest

IF is an unsupervised anomaly detection method that isolates anomalies by randomly selecting features and split values, based on the principle that anomalies are few and different, requiring fewer random splits to isolate compared to normal observations (17). See supplementary section (A.1) for the mathematical foundation of the IF algorithm.

Separate models were trained for each pregnancy outcome group (i.e. livebirth term, livebirth preterm, stillbirth or miscarriage) due to the varying nature of GA distribution across these outcomes.

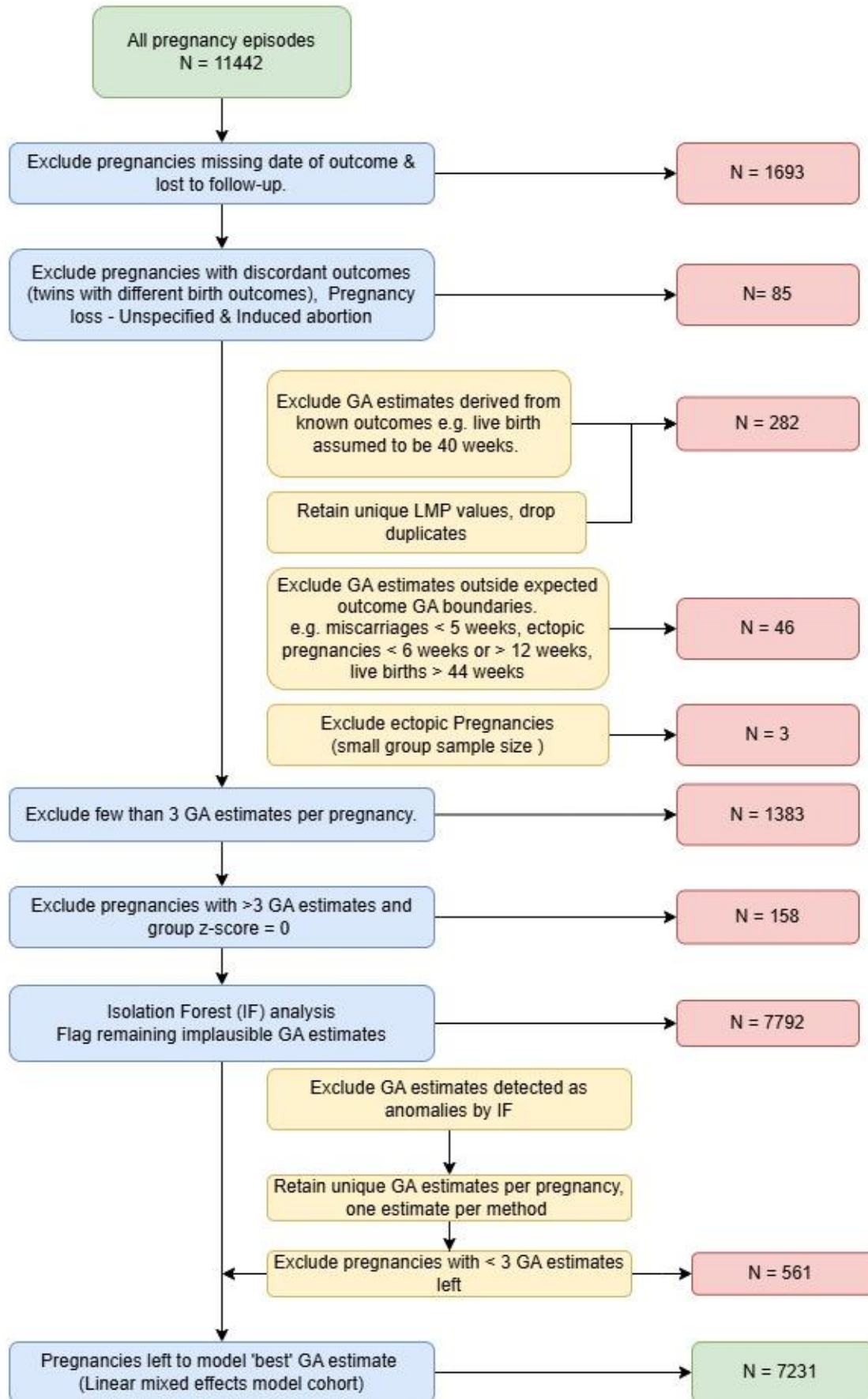


Figure 1: Exclusion Criteria Diagram

A grid search was conducted to determine the optimal hyperparameters for each outcome group. These parameters were 'contamination' (v), 'n_estimators' (T) and 'max_samples' (ψ). The (v) parameter, represents the expected proportion of anomalies, and directly influences the decision threshold. The algorithm uses this value to set the threshold on the anomaly score. The top (v) percent of points with the highest anomaly scores (shortest path lengths) are classified as anomalies. The (T) parameter controls the number of individual decision trees to build in the forest. Each tree is trained on a random subset of the data. An anomaly score is calculated for each data point by averaging the path lengths (number of splits required to isolate it) across all trees. Anomalies will have shorter average path lengths. The (ψ) parameter represent the number of samples used to train each individual tree. Setting it to "auto" typically defaults to a minimum of 256 samples.

The optimisation criterion was the silhouette score, which quantifies the separation between the resulting anomaly and normal clusters (24). The parameter combination yielding the highest score was selected for the final model (See **Table 1**).

Table 1: Isolation Forest hyperparameters and silhouette scores by pregnancy outcome

Outcome	Contamination (v)	Estimators (T)	Max samples (ψ)	CV silhouette	Tuning silhouette
Live birth term	0.005	100	512	0.638 ± 0.030	0.631
Live birth preterm	0.020	300	512	0.411 ± 0.014	0.412
Stillbirth	0.030	300	512	0.404 ± 0.066	0.447
Miscarriage	0.050	100	256	0.318 ± 0.017	0.356

CV = cross-validation; v = contamination parameter; T = number of estimators (trees); ψ = maximum number of samples used to train each tree.

CV silhouette: mean $\pm$ SD silhouette coefficient across cross-validation folds. Tuning silhouette: silhouette coefficient of the final tuned model.

Silhouette coefficients range from -1 to 1 ; higher values indicate better-defined cluster separation between normal and anomalous observations.

The feature set for each outcome was tailored to its clinical context (see **Table 2**). The raw dataset included prenatal (LMP, Ultrasound, fundal height, GA at subsequent ANC visits) and postnatal (Foot length, Ballard score) GA estimates. For each record, the following additional variables were captured: the method of GA measurement, expected date of delivery (EDD), actual delivery date, pregnancy outcome, the clinically expected GA range for that outcome, and a unique pregnancy identifier. We engineered the following features for each GA measurement to prepare the data for anomaly detection (See **Table 2**). The EDD related feature was excluded for pregnancy loss outcomes due to reduced reliability for pregnancies not reaching term. All continuous features were standardised using z-score normalisation.

A 5-fold cross-validation scheme was employed to provide a robust estimate of generalisability and to guard against overfitting. The model was trained on four folds and its performance evaluated on the held-out fold using the silhouette score. This process was repeated five times: the mean silhouette score and its standard deviation are reported.

We conducted feature importance analysis using two complementary approaches to understand which features were most influential in anomaly detection across different pregnancy outcomes.

(a) We calculated standardised effect sizes to measure how much anomalous cases differed from normal cases for each GA feature. This approach identified features with the largest differences between the two groups, providing insights into the characteristics that distinguish anomalies.

(b) We assessed each feature's importance to the anomaly detection model by repeatedly shuffling its values and measuring how much this degraded model performance. Features causing significant performance drops when randomised were considered critical to the model's decision-making process.

Table 2: Outcome-specific feature sets and their derivation

Feature	Description
Gestational age at expected delivery date (GAedd)	Calculated for each method using its estimated start date and the expected date of delivery.
Gestational age at delivery (GAdel)	Calculated for each method using its estimated start date and the actual delivery date.
GA z-score across all pregnancies per outcome (Ozscore)	Each GA at delivery was standardised into a z-score relative to the overall population distribution for that specific pregnancy outcome.
GA deviation from expected range (GAdeviation)	The absolute difference between the calculated GA at delivery and the (a) minimum and (b) maximum of the outcome-specific expected GA range.
Z-score for gestation within a pregnancy (Zga)	The mean gestation was computed and then a z-score was calculated for each measurement within its pregnancy. Measurements with a z-score of zero were excluded as non-informative.
Expected maximum GA per outcome (EGA Max)	Ectopic: 12 weeks; Miscarriage: 27.9 weeks; Preterm live birth: 36.9 weeks; Term live birth: 44 weeks; Stillbirth: 44 weeks.
Expected minimum GA per outcome (EGA Min)	Ectopic: 6 weeks; Miscarriage: 5 weeks; Preterm live birth: 28 weeks; Term live birth: 37 weeks; Stillbirth: 28 weeks.

GA = gestational age; EDD = estimated date of delivery; EGA = expected gestational age.
All gestational age values are expressed in completed weeks.

We conducted a manual review of the detected cases to iteratively refine the GA anomaly detection algorithm. This process verified the flagged cases represented genuine errors. We used a stratified random sample, with 50% representation from each outcome category. A quality assurance officer and a study clinician performed the review, cross-referencing each flagged estimate against other available GA measurements and the recorded pregnancy outcomes. All

anomalies were profiled by estimation method, and ultrasound-related errors were corrected using the available image scans where possible.

2.2.2 Stage 2: Linear mixed-effects model for GA estimation

We used linear mixed-effects models to combine multiple GA measurements per pregnancy into a single, calibrated estimate of GA at the time of the actual pregnancy outcome. One date per estimation method was retained for each pregnancy, including only pregnancies with at least three distinct estimation methods. Unique multiple estimates from GA at subsequent ANC visits were included for each pregnancy. After these exclusions (see **Figure 1**), the final analytic dataset was derived from pregnancies screened for anomalies in the previous stage. The analytic dataset contained multiple GA measurements for each pregnancy, derived from different estimation methods including ultrasound, LMP, fundal height, Ballard Score, and foot length. Each pregnancy had an observed birth outcome (Live Birth Term, Live Birth Preterm, Stillbirth, or Miscarriage) and an associated birth outcome date, which we treated as the event date for GA at outcome.

Let GA_{ij} denote the GA (in days) for pregnancy i as estimated by method j at the pregnancy birth outcome date. We specified the following linear mixed-effects model:

$$GA_{ij} = \beta_0 + \beta_{GA_methodj} + \beta_{outcomek} + u_i + \varepsilon_{ij},$$

where β_0 is the intercept corresponding to the first trimester ultrasound GA measurement and term birth outcome category, $\beta_{GA_methodj}$ are fixed effects representing systematic differences (biases) of each GA estimation method relative to first-trimester ultrasound measurement, and $\beta_{outcomek}$ are fixed effects for pregnancy outcome categories (Live Birth Term, Live Birth Preterm, Stillbirth, Miscarriage). The term $u_i \sim N(0, \sigma_u^2)$ is a random intercept for pregnancy i , capturing pregnancy-specific deviation from the population mean after accounting for method and outcome, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the residual error. In R notation, the model is:

$$GA_birth \sim GA_method + birth_outcome + (1 | pregnancy_id)$$

This formulation allows the fixed effects for GA_method to calibrate all methods onto the scale of the first-trimester ultrasound method. The random intercept u_i then synthesises multiple GA measurements within each pregnancy into a single pregnancy-specific deviation, borrowing strength across methods and shrinking individual estimates toward the population mean when data are sparse. In practical terms, this means that even for a woman with only one or two available GA measurements, the model draws on patterns observed across all pregnancies in the dataset to produce a more reliable individual conception date estimate than any single method alone could provide.

For each pregnancy, we derived a model-based ‘best’ GA at outcome by combining the fixed-effect prediction (for the chosen reference GA method and the pregnancy’s actual outcome category) with the BLUP estimate of the random intercept u_i . This yielded one

calibrated GA estimate per pregnancy on the reference method scale, applicable to pregnancies with and without early ultrasound data. Conception date was then approximated for each pregnancy by subtracting this best GA (in days) from the observed birth outcome date. These pregnancy-specific GA and conception estimates were used in subsequent analyses such as data linkage for pregnancy episodes, classification of the timing of medication exposures by trimester, and for method comparison bias summaries.

2.3 Implementation and Computational Details

All analyses were conducted in R version 4.4.3 (25) and Python 3.11 (26). The Isolation Forest anomaly detection was implemented using scikit-learn (version 1.0.2) (27). The Python and R code are available at GitHub (28).

3 Results

3.1 Dataset Characteristics

Of the 11,442 pregnancies identified from a cohort of 35,953 enrolled women, 7,792 met the inclusion criteria and contributed 53,926 GA observations to the analysis. Table 3 presents distribution of GA dating methods by outcome.

Table 3: Distribution of gestational age methods by pregnancy outcome

GA dating method	Overall	Live birth preterm	Live birth term	Miscarriage	Stillbirth
Total pregnancies, n	7,792	382	7,162	106	142
Fundal height (FH)	1,458 (18.7%)	46	1,379	4	29
GA at ANC ^a	6,451 (82.8%)	318	5,909	103	121
Ballard score (BS)	5,286 (67.8%)	168	5,118	0	0
Foot length (FL)	4,142 (53.2%)	231	3,911	0	0
First-trimester US	957 (12.3%)	26	876	43	12
Second-trimester US	3,530 (45.3%)	122	3,328	9	71
Third-trimester US	1,764 (22.6%)	55	1,674	0	35
Mid-month LMP	742 (9.5%)	30	680	13	19
Self-reported LMP	6,764 (86.8%)	305	6,217	106	136

GA = gestational age; ANC = antenatal care; LMP = last menstrual period.

^a GA at ANC: date computed from progressive gestation recorded at subsequent antenatal care visits.

US: Ultrasound scans.

Counts represent at least one documentation of a source per pregnancy. Percentages in the Overall column are of total pregnancies.

3.2 Anomaly Detection Results

Anomaly rates varied across pregnancy outcomes, with Live Birth Term having the lowest rate (0.5%, 253/50,517), followed by Live Birth Preterm (2.0%, 43/2,118), Stillbirth (2.8%, 25/878),

and Miscarriage having the highest anomaly rate (5.1%, 21/413). Anomaly rates were consistently higher in pregnancies with adverse outcomes.

Anomaly distribution varied by pregnancy outcome and GA method (**Table 4**). GA estimated at subsequent antenatal clinic visits was the predominant source of anomalies, accounting for nearly half of all detected anomalies (160/342, 46.8%) and contributing the highest proportion of anomalies across all pregnancy outcomes. LMP and ultrasound estimates were the second and third largest sources, contributing 74 (21.6%) and 42 (12.3%) anomalies, respectively. Among clinical measurement methods, Ballard score and foot length, contributed anomalies in live birth outcomes, as these are only measured in livebirths.

IF model feature performance is detailed in the supplementary section **A.2**.

Table 4: Distribution of anomalous gestational age estimates by dating source and pregnancy outcome

GA dating source	Live birth preterm	Live birth term	Stillbirth	Miscarriage	Total n (%)
GA at ANC visit	18	120	11	11	160 (46.8%)
Self-reported LMP	8	55	5	6	74 (21.6%)
Ultrasound scan	1	30	7	4	42 (12.3%)
Ballard score	10	18	0	0	28 (8.2%)
Fundal height	1	18	2	0	21 (6.1%)
Foot length	4	10	0	0	14 (4.1%)
Mid-month/year of LMP	1	2	0	0	3 (0.9%)
Total	43	253	25	21	342 (100%)

GA = gestational age; ANC = antenatal care; LMP = last menstrual period.

Anomalies identified by Isolation Forest algorithm. Values represent counts of flagged estimates per method.

Total column shows n (%) of all anomalies across outcomes attributable to each source.

3.3 Linear mixed effects results

Table 5 summarises agreement between each GA assessment method and the mixed-effects, first-trimester ultrasound calibrated estimate (mean GA: 278.6 days at livebirth term). The mixed-effects model revealed differences in precision across dating methods, with a clear systematic shift observed for ultrasound assessments performed later in gestation. Although ultrasound retained high individual-level precision (SD: 7.5–8.9 days), mean bias increased with advancing gestation, from 3.15 days in the second trimester to 6.41 days in the third. Despite this temporal variation, ultrasound-based estimates remained more reliable for individual GA assessment than alternative methods. In contrast, the Ballard score, foot length, fundal height, and last menstrual period (LMP) exhibited minimal mean bias (< 2 days) but were characterised by substantial individual-level uncertainty, with variability exceeding 35 days. This wide dispersion indicates that, while these methods may appear accurate at the population level, they introduce considerable noise when estimating GA for individual pregnancies.

Table 5: Method-specific bias in gestational age relative to first-trimester ultrasound scan, sorted by absolute bias

GA method	Mean bias $\pm$ SD (days)	95% Limits of Agreement (weeks)	n
US1	0.00 $\pm$ 8.12	-15.90 to 15.90	939
US2	3.15 $\pm$ 7.45	-11.50 to 17.80	3,431
US3	6.41 $\pm$ 8.88	-11.00 to 23.80	1,694
FH	1.37 $\pm$ 11.10	-20.30 to 23.00	1,424
BS	-0.03 $\pm$ 9.23	-18.10 to 18.10	5,161
LMP	-0.82 $\pm$ 10.10	-20.70 to 19.00	6,319
ANC_GA	-2.71 $\pm$ 11.20	-24.70 to 19.30	6,050
M_LMP	0.05 $\pm$ 11.80	-23.00 to 23.10	737
FL	-3.92 $\pm$ 9.38	-22.30 to 14.50	4,101

GA = gestational age; SD = standard deviation; LOA = limits of agreement.

US1 = first-trimester ultrasound; US2 = second-trimester ultrasound; US3 = third-trimester ultrasound;

ANC_GA = antenatal care gestational age estimate; FH = fundal height; BS = Ballard score;

FL = foot length; M_LMP = maternally recalled last menstrual period; LMP = last menstrual period (self-reported).

Mean bias = mean difference relative to US1 (reference); positive values indicate overestimation.

95% LOA calculated as mean bias $\pm$ 1.96 $\times$ SD (Bland–Altman method).

3.3.1 Example pregnancies illustrating method-specific and model-based conception dates

Figure 2 illustrates, for three example pregnancies (p1, p2, p3), how multiple method-specific GA measurements are combined into a single model-based conception date. The blue points represent conception dates implied by individual methods, including fundal height (FH), LMP, and postnatal assessments (BS, FL), highlighting the inherent variability across methods. The red diamonds and vertical red lines denote the model ‘best’ estimate, representing the weighted estimate that incorporates pregnancy-specific random effects to calibrate these disparate inputs to a first-trimester ultrasound scale. The numeric labels, indicate the ‘shrinkage’ in days, representing the distance the model pulls each method toward the consensus to correct for systematic bias and individual noise. The figure demonstrates that a consistent predicted conception date can be obtained even for pregnancies with diverse combinations of measurement types, thereby providing a harmonised GA metric for use in subsequent analyses.

Linear Mixed Effects Model: Shrinkage of Dating Methods toward Consensus

The vertical line represents the model best estimate.

Days labels represent the distance from the method-specific date to model estimate.

Legend : • Method-Specific Date ♦ Model Consensus (Best)

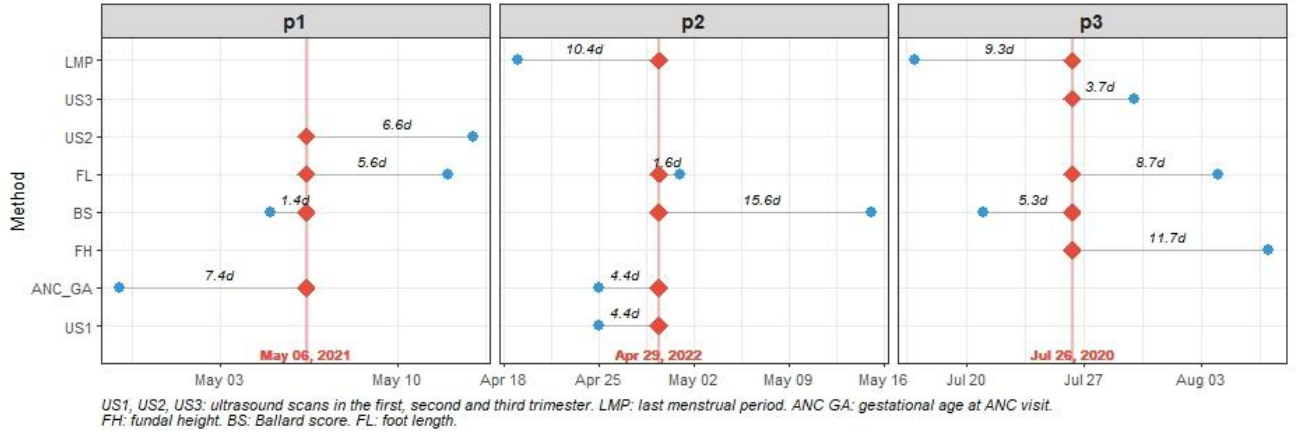


Figure 2: Conception date estimates for three randomly selected example pregnancies (p1, p2, p3), illustrating method-specific (blue) and model-based (red) dates. The red diamond represents the BLUP-derived conception date calibrated to the first-trimester ultrasound scale, combining fixed-effect predictions with the pregnancy-specific random intercept. Distances (d) indicate the number of days between each method-specific estimate and the model consensus date.

4 Discussion

This study demonstrates that disparate GA measurements from routine care datasets can be synthesised into calibrated ‘best estimate’ conception dates by using a two-stage approach combining machine learning-driven GA anomaly detection with linear mixed effects modelling. The IF algorithm revealed that data quality was not uniformly distributed across pregnancy outcomes; anomaly rates increased alongside pregnancy complications, 5.1% for miscarriages compared to only 0.5% for term live births. These discrepancies were largely attributed to data entry errors in gestational estimates in the antenatal clinic records and LMP recall. Ultrasound measurements had the highest individual-level precision (SD: 7.5–8.9 days) and progressive mean bias with advancing gestation (3–6 days). Ballard score, fundal height, and LMP had negligible bias (< 2 days) at the population level with high uncertainty at the individual level (spread > 35 days). Foot length underestimated GA by approximately 4 days. The model generated a de-biased “best” GA estimate calibrated to the first-trimester ultrasound scale that accounts for both systematic bias and individual variation even in pregnancies with limited ultrasound scan and varying GA methods. Importantly, because the model pools information across the entire dataset, it generates a more reliable conception date for each individual woman than her own measurements alone could provide. This individual-level calibration is the key practical advantage of the approach, enabling accurate classification of medication exposures by trimester and supporting downstream pharmacovigilance analyses at the individual level.

These findings mirror broader methodological challenges established in obstetric epidemiology. The varying data quality across pregnancy outcomes is consistent with literature

linking adverse pregnancy outcomes to irregular dating (29,30). The progressive bias observed in ultrasound aligns with established foetal growth standards, where biological variability in foetal size becomes less closely related to true gestational age, especially in the second and third trimesters (31). The contrast between the population-level accuracy and individual-level imprecision of clinical methods reflects a common phenomenon where population averages mask significant individual dating errors (32). Similarly, the underestimation observed in foot length aligns with recent systematic reviews indicating significant population heterogeneity (33) and nutritional stunting (34) limits its utility as a continuous estimator without specific calibration. The critical need for such calibration is highlighted by findings from Moore et al (35), which demonstrated that uncorrected metrics often may overestimate GA in preterm infants, thereby obscuring high-risk cohorts. By addressing these established pitfalls, our model reconciles these divergent metrics to generate a de-biased “best” GA estimate. These findings support the use of the model for pregnancy episode record linkage and downstream pharmacovigilance analyses.

Digital routine care datasets constitute a robust resource for perinatal research, capturing comprehensive longitudinal details ranging from triage and laboratory diagnostics to disease management and therapeutic interventions (36). However, the effective utilisation of these data is often compromised by limited system interoperability, which results in fragmented or duplicative entries for the same patient across different facilities and electronic platforms (37). This fragmentation is problematic for GA estimation, where a single pregnancy typically accumulates multiple, conflicting dating records throughout the gestation period (4). In low-resource settings, this challenge is compounded by the scarcity of first-trimester ultrasound, as most women initiate antenatal care (ANC) during the second or third trimester (10). As a result, one pregnancy record can contain multiple, conflicting GA estimates, and without a first-trimester gold standard, it is difficult to determine which date is most accurate. Such inconsistencies can introduce systematic bias into pregnancy outcome classifications, particularly inflating rates of preterm birth or misidentifying growth-restricted infants as constitutionally small (38). Despite these structural limitations, these datasets reflect authentic clinical encounters and offer the large-scale, real-world evidence necessary for population-level perinatal research (39). Recognising and addressing these data quality issues is central to unlocking the value of routine care records, particularly where early ultrasound is unavailable.

Data-driven innovations offer a pathway to mitigate these quality issues and enhance the analytic value of routine health records. Unsupervised learning techniques, such as the anomaly detection (IF) employed in this study, can identify implausible entries, thereby improving internal data consistency without discarding vast amounts of clinical information (17,40). When paired with statistical methods like linear mixed-effects models, diverse dating sources can be integrated into unified, bias-adjusted GA estimates (35,41). Emerging technologies such as digital health platforms are improving the temporal resolution and accuracy of data capture at the point of care (42). Over time, combining machine learning algorithms with digital record systems can enable adaptive, continuously improving models for GA estimation and perinatal

monitoring. Together, these innovations have the potential to transform routine care data from fragmented clinical records into reliable, research-grade resources, capable of supporting both individual clinical care and large-scale perinatal research.

Precise GA estimation is a prerequisite for the accurate classification of preterm births and small-for-gestational-age (SGA) infants, which in turn drives critical pharmacovigilance outcomes and neonatal survival strategies (43,44). Adopting data-driven methods for perinatal research in low-resource settings strengthens the evidence base for the better utilisation of routine digital health records (39,45). However, as currently implemented, this methodology primarily serves retrospective research purposes. Translating these models to directly benefit healthcare providers requires integrating them into existing, interoperable digital health platforms. The necessary next step is the development of a user-friendly clinical decision support interface embedded within routine Electronic Medical Records (EMRs). In practice, a clinician would simply enter standard clinical parameters into the EMR and click a button to run the model in the background, prompting the system to instantly compute and return an optimised GA estimate. This seamless integration of ML tools into routine care workflows offers real-time decision support for clinicians, enabling timely interventions that might otherwise be missed due to data ambiguity (46). To achieve this, there must be a commitment to investment in data quality governance, training, and system interoperability to ensure equitable adoption and sustainable progress in low-resource settings. The convergence of data science and maternal health systems positions emerging technologies to standardise GA estimation, optimise pregnancy outcome tracking, and strengthen research capacity across diverse, resource-limited contexts.

4.1 Limitations

Three limitations warrant consideration. First, the linear mixed-effects framework did not explicitly model method-specific residual variance or complex correlation structures between clinical assessments. For instance, if diverse dating methods (e.g., fundal height and LMP) share systematic errors due to provider dependence, our model may have underestimated the remaining measurement error for individual assessments. Future work should explore more flexible approaches, including Bayesian hierarchical models and precision-weighted estimation strategies, to better propagate this uncertainty into downstream inferences on pregnancy outcomes. Second, the scarcity of early ultrasound scans (939 (12.9%)), our calibration ‘gold standard’, relative to late scans may compromise the stability of the estimates. This imbalance is particularly pronounced in rare outcome groups like miscarriages, potentially limiting the internal validity of the calibration for these high-risk cohorts. Third, the generalisability of our findings may be constrained by the context of the data source. Because this analysis was conducted within the MiMBa Pregnancy Exposure Registry, the data completeness and clinical workflows may reflect a higher standard of monitoring and active surveillance than typical routine healthcare facilities. Applying this model to other low-resource settings with different demographic profiles, data collection protocols, or less robust health information systems will

require external validation. Ultimately, larger datasets with greater availability of first-trimester ultrasound and multi-site clinical data will be essential to refine these data-driven approaches and to determine how best to calibrate routine care-based estimates in diverse, real-world settings.

Step-by-Step Implementation for GA Estimation in Routine Datasets

Research teams seeking to apply this methodology to multi-register routine healthcare datasets may find the following workflow useful:

1. **Data consolidation:** Extract and link pregnancy episode records across available facility registers (e.g., ANC, maternity, EMRs) to capture all available GA dating methods (e.g., LMP, fundal height, ultrasound scans, postnatal assessments).
2. **Anomaly detection:** Apply programmatic data-cleaning rules to identify and filter out implausible or physiologically impossible clinical observations before modelling.
3. **Statistical integration:** Employ a systematic modelling approach, such as a linear mixed-effects, to combine multiple conflicting but plausible GA measurements into a single estimate, explicitly accounting for pregnancy-level variability and method-specific biases.
4. **Standard calibration:** Calibrate the model outputs against a highly reliable subset of your data (e.g., available first-trimester ultrasound scans) to anchor the estimates to a clinical gold standard.
5. **Output generation:** Generate a unified, optimised ‘best GA’ metric for each pregnancy to be used in downstream epidemiological classifications (e.g., preterm birth, SGA) and pharmacovigilance tracking.

Conclusion

Even in the absence of widespread first-trimester ultrasound, routine care datasets can be leveraged to generate more accurate GA estimates when combined with robust data-cleaning and modelling strategies. This approach maximises the informational value of imperfect routine data and supports more reliable classification of pregnancy outcomes in low-resource settings. By pooling information across all pregnancies, the model generates a calibrated, individual-level conception date for each woman, even when her own measurements are limited, a critical advantage for pharmacovigilance analyses requiring accurate individual pregnancy timing. These methods are particularly well-suited to pharmacovigilance registries and large-scale observational studies, where the data quality controls available in clinical trials are rarely feasible. However, translating this approach into real-time clinical practice will require embedding these algorithms into interoperable electronic medical records to provide point-of-care decision support for frontline clinicians.

5 List of abbreviations

GA – Gestational age
LMP – Last menstrual period
SD – Standard deviation
LMICs - Low- and middle-income countries
ANC - Antenatal care
IF – Isolation Forest
ML – Machine learning
BLUP - Empirical best linear unbiased prediction
EDD – Expected date of delivery
EGA – Expected gestational age
FH – Fundal height
BS – Ballard score
US – Ultrasound
FL – Foot length
LOA – Limits of agreement
SGA – Small for gestation age
EMRs – Electronic medical records systems

6 Declarations

Ethics approval and consent to participate

The primary study received ethical approval from the LSTM Research Ethics Committee (reference: 2056) and KEMRI SERU (reference: 4093). The current study is a secondary analysis of data collected under that approval; no additional ethical review was required, as secondary analyses of this dataset were permitted under the terms of the original approval. All participants in the primary study provided written informed consent prior to data collection. Data cleaning and analysis were conducted by the same study team responsible for the primary data collection.

Consent for publication

Not applicable

Availability of data and materials

The datasets used for analysis are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was undertaken under the MiMBa Pregnancy Exposure Registry project (<https://clinicaltrials.gov/study/NCT04825782>) updated on 2021-04-01, sponsored by Medicine for Malaria Venture - <https://www.mmv.org/> awarded to FtK and SD.

Award number: PO19/01469

The funders did not play a role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Clinical trial number

Not applicable

Authors' contributions

MC conceptualised and executed the methodical approach under the close supervision of ML. SD and ML reviewed the results. MC drafted the manuscript, which was subsequently reviewed by ML, SD, MT, SK, MO, BO, GO, and FtK. KM, MA, and MC performed manual reviews of flagged records and corrected ultrasound data. BA and TO conducted data cleaning. SD, FtK, HB, and SK serve as the principal investigators of the MiMBa Pregnancy Exposure Registry (primary study).

Acknowledgements

The authors would like to thank all the women who participated in the primary study; their willingness to contribute their time and data made this research possible. We acknowledge the support and assistance provided by the MiMBa study research team, Quantitative Engineering Design (QED.ai) management and technical teams, Homabay county department of health, county health records information management and sub-county teams.

References

1. American College of Obstetricians and Gynecologists. Committee Opinion No. 700: Methods for Estimating the Due Date. *Obstetrics & Gynecology*. 2017;129: e150–e154.
2. Rappazzo KM, Lobdell DT, Messer LC, Poole C, Daniels JL. The comparison of gestational dating methods and implications for exposure-outcome associations: an example with PM2.5 and preterm birth. *Occupational and environmental medicine*. 2017;74(2): 138–143. <https://doi.org/10.1136/oemed-2016-103833>.
3. *WHO recommendations on antenatal care for a positive pregnancy experience*. <https://www.who.int/publications/i/item/9789241549912> [Accessed 1st April 2026].

4. Lynch CD, Zhang J. The research implications of the selection of a gestational age estimation method. *Paediatric and Perinatal Epidemiology*. 2007;21(Suppl 2): 86–96. <https://doi.org/10.1111/j.1365-3016.2007.00865.x>.
5. Skinner C, Mount CA. Sonography Assessment of Gestational Age. In: *StatPearls. Treasure Island (FL): StatPearls Publishing; 2025*. <http://www.ncbi.nlm.nih.gov/books/NBK570610/> [Accessed 2nd September 2025].
6. Alyahyaw A, Adam GK, AlHabardi N, Adam I. Problems with gestational age estimation by last menstrual period and ultrasound among late antenatal care attendant women in a low-resource setting in Africa, Sudan. *Journal of Ultrasound*. 2024;27(1): 129–135. <https://doi.org/10.1007/s40477-023-00844-x>.
7. Whelan R, Schaeffer L, Olson I, Folger LV, Alam S, Ajaz N, et al. Measurement of symphysis fundal height for gestational age estimation in low-to-middle-income countries: A systematic review and meta-analysis. *PLoS ONE*. 2022;17(8): e0272718. <https://doi.org/10.1371/journal.pone.0272718>.
8. Ballard JL, Khoury JC, Wedig K, Wang L, Eilers-Walsman BL, Lipp R. New Ballard Score, expanded to include extremely premature infants. *The Journal of Pediatrics*. 1991;119(3): 417–423. [https://doi.org/10.1016/s0022-3476\(05\)82056-6](https://doi.org/10.1016/s0022-3476(05)82056-6).
9. Unger H, Thriemer K, Ley B, Tinto H, Traoré M, Valea I, et al. The assessment of gestational age: a comparison of different methods from a malaria pregnancy cohort in sub-Saharan Africa. *BMC Pregnancy and Childbirth*. 2019;19: 12. <https://doi.org/10.1186/s12884-018-2128-z>.
10. Kase BF, Seifu BL, Mare KU, Shibeshi AH, Asebe HA, Gemedo K, et al. Time to first antenatal care visit and its predictors among women in Kenya: Weibull gamma shared frailty model (based on the recent 2022 KDHS data). *BMC Pregnancy and Childbirth*. 2025;25: 50. <https://doi.org/10.1186/s12884-025-07178-y>.
11. *Kenya Demographic and Health Survey - 2022 - Kenya National Bureau of Statistics*. <https://www.knbs.or.ke/reports/kdhs-2022/> [Accessed 2nd May 2026].
12. Stewart KA, Navarro SM, Kambala S, Tan G, Poondla R, Lederman S, et al. Trends in Ultrasound Use in Low and Middle Income Countries: A Systematic Review. *International Journal of MCH and AIDS*. 2020;9(1): 103–120. <https://doi.org/10.21106/ijma.294>.
13. Hanson C, Nyaga L, Leekha N, Mantel M, Kedenge S, Gitonga CW, et al. Routine antenatal ultrasound by nurse-midwives in rural Kenya: a pragmatic trial assessing feasibility and effects of the Mimba Yangu (My Pregnancy) project. *Frontiers in Global Women's Health*. 2025;6: 1555547. <https://doi.org/10.3389/fgwh.2025.1555547>.
14. Couper I, Ray S, Blaauw D, Rugarabamu P, Cameron D, Covin G, et al. Curriculum and training needs of mid-level health workers in Africa: a situational review from Kenya, Nigeria, South Africa and Uganda. *BMC Health Services Research*. 2018;18: 553. <https://doi.org/10.1186/s12913-018-3362-9>.
15. Agoro OO, Osga BO, Adoyo M. Supportive supervision for medicines management in government health facilities in Kiambu County, Kenya: a health workers' perspective. *Pan African Medical Journal*. 2015;20. <https://doi.org/10.11604/pamj.2015.20.237.5872>.

16. Shamba D, Day LT, Zaman SB, Sunny AK, Tarimo MN, Peven K, et al. Barriers and enablers to routine register data collection for newborns and mothers: EN-BIRTH multi-country validation study. *BMC Pregnancy and Childbirth*. 2021;21(Suppl 1): 233. <https://doi.org/10.1186/s12884-020-03517-3>.
17. Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE; 2008. p. 413–422. <https://doi.org/10.1109/ICDM.2008.17>.
18. Preez A du, Bhattacharya S, Beling P, Bowen E. Fraud detection in healthcare claims using machine learning: A systematic review. *Artificial Intelligence in Medicine*. 2025;160: 103061. <https://doi.org/10.1016/j.artmed.2024.103061>.
19. Bukhari A, Alshemaimri B, Abbas T, Alharbey R, Aloufi K, Daud A. Anomaly-based threat detection in smart health using machine learning. *BMC Medical Informatics and Decision Making*. 2024;24(1): 347. <https://doi.org/10.1186/s12911-024-02760-4>.
20. Penev YP, Buchanan TR, Ruppert MM, Liu M, Shekouhi R, Guan Z, et al. Electronic health record data quality and performance assessments: scoping review. *JMIR Medical Informatics*. 2024;12: e58130. <https://doi.org/10.2196/58130>.
21. Fraser HS, Mugisha M, Bacher I, Ngenzi JL, Seebregts C, Umubyeyi A, et al. Factors influencing data quality in electronic health record systems in 50 health facilities in Rwanda and the role of clinical alerts: cross-sectional observational study. *JMIR Public Health and Surveillance*. 2024;10: e49127. <https://doi.org/10.2196/49127>.
22. *Malaria in Mothers and Babies Pregnancy Exposure Registry - MiMBa*. LSTM. <https://www.lstmed.ac.uk/MiMBa> [Accessed 1st April 2026].
23. *ScanForm*. ScanForm. <https://about.scanform.ai> [Accessed 30th April 2026].
24. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987;20: 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
25. *R: The R Project for Statistical Computing*. <https://www.r-project.org/> [Accessed 1st April 2026].
26. *Download Python*. Python.org. <https://www.python.org/downloads/> [Accessed 1st April 2026].
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85): 2825–2830.
28. mchepkirui. *mchepkirui/Gestation-Age-Estimation-Isolation-Forest-and-linear-mixed-effect-model*. 2026. <https://github.com/mchepkirui/Gestation-Age-Estimation-Isolation-Forest-and-linear-mixed-effect-model> [Accessed 29th April 2026].
29. Sundermann AC, Jasper EA, Kumar SE, Hartmann KE, Velez Edwards DR. Dating Discrepancies on Research Ultrasonography and Risk of Pregnancy Loss in a Prospective

Cohort. *Obstetrics & Gynecology*. 2025;146(4): 515–523. <https://doi.org/10.1097/AOG.0000000000006031>.

30. Rashid H, Ma E, Ferdous F, Ekström EC, Wagatsuma Y. First-trimester fetal growth restriction and the occurrence of miscarriage in rural Bangladesh: A prospective cohort study. *PLoS ONE*. 2017;12(7): e0181967. <https://doi.org/10.1371/journal.pone.0181967>.
31. Papageorgiou AT, Kemp B, Stones W, Ohuma EO, Kennedy SH, Purwar M, et al. Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound in Obstetrics & Gynecology*. 2016;48(6): 719–726. <https://doi.org/10.1002/uog.15894>.
32. Macaulay S, Buchmann EJ, Dunger DB, Norris SA. Reliability and validity of last menstrual period for gestational age estimation in a low-to-middle-income setting. *Journal of Obstetrics and Gynaecology Research*. 2019;45(1): 217–225. <https://doi.org/10.1111/jog.13801>.
33. Tikmani SS, Mårtensson T, Khalid S, Uzair M, Ali Q, Rahim A, et al. Assessing the diagnostic accuracy of postnatal clinical scoring methods and foot length measurement for estimating gestational age and birthweight of newborns in low- and middle-income countries: a systematic review and meta-analysis. *BMJ Paediatrics Open*. 2024;8(1): e002717. <https://doi.org/10.1136/bmjpo-2024-002717>.
34. Mengi A, Vallely LM, Laman M, Jally E, Kulimbao J, Warel S, et al. The use of newborn foot length to identify low birth weight and preterm babies in Papua New Guinea: A diagnostic accuracy study. *PLOS Global Public Health*. 2023;3(6): e0001924. <https://doi.org/10.1371/journal.pgph.0001924>.
35. Moore KA, Simpson JA, Thomas KH, Rijken MJ, White LJ, Dwell SLM, et al. Estimating Gestational Age in Late Presenters to Antenatal Care in a Resource-Limited Setting on the Thai-Myanmar Border. *PLOS ONE*. 2015;10(6): e0131025. <https://doi.org/10.1371/journal.pone.0131025>.
36. Smeets HM, Kortekaas MF, Rutten FH, Bots ML, Van Der Kraan W, Daggelders G, et al. Routine primary care data for scientific research, quality of care programs and educational purposes: the Julius General Practitioners' Network (JGPN). *BMC Health Services Research*. 2018;18(1): 735. <https://doi.org/10.1186/s12913-018-3528-5>.
37. Chepkirui M, Dellicour S, Musuva R, Odero I, Omondi B, Omondi B, et al. Missed opportunities for digital health data use in healthcare decision-making: A cross-sectional digital health landscape assessment in Homa Bay county, Kenya. *PLOS Digital Health*. 2025;4(6): e0000870. <https://doi.org/10.1371/journal.pdig.0000870>.
38. Vitral GLN, Romanelli RM de C, Leonel TA, Souza Gaspar J de, Aguiar RALP de, Reis ZSN. Influence of different methods for calculating gestational age at birth on prematurity and small for gestational age proportions: a systematic review with meta-analysis. *BMC Pregnancy and Childbirth*. 2023;23(1): 106. <https://doi.org/10.1186/s12884-023-05411-0>.
39. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. *BMC Health Services Research*. 2020;20(1): 790. <https://doi.org/10.1186/s12913-020-05660-1>.

40. Röchner P, Rothlauf F. Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. *BMC Medical Research Methodology*. 2023;23(1): 125. <https://doi.org/10.1186/s12874-023-01946-0>.
41. Stirnemann J, Villar J, Salomon LJ, Ohuma E, Ruyan P, Victora CG, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*. 2014;384(9946): 869–879. [https://doi.org/10.1016/S0140-6736\(14\)61490-2](https://doi.org/10.1016/S0140-6736(14)61490-2).
42. Labrique A, Vasudevan L, Kochi E, Fabricant R, Mehl G. Digital health and health systems of the future. *Global Health: Science and Practice*. 2018;6(Suppl 1): S1–S4. <https://doi.org/10.9745/GHSP-D-18-00342>.
43. Schlaudecker EP, Munoz FM, Bardají A, Boghossian NS, Khalil A, Mousa H, et al. Small for gestational age: Case definition & guidelines for data collection, analysis, and presentation of maternal immunisation safety data. *Vaccine*. 2017;35(48, Part A): 6518–6528. <https://doi.org/10.1016/j.vaccine.2017.01.040>.
44. Quinn JA, Munoz FM, Gonik B, Frau L, Cutland C, Mallett-Moore T, et al. Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine*. 2016;34(49): 6047–6056. <https://doi.org/10.1016/j.vaccine.2016.03.045>.
45. Todd J, Mahande MJ. Editorial: The Use of Routine Health Data in Low- and Middle-Income Countries. *Frontiers in Public Health*. 2020;8. <https://doi.org/10.3389/fpubh.2020.00413>.
46. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 2019;25(1): 44–56. <https://doi.org/10.1038/s41591-018-0300-7>.
47. Breiman L. Random Forests. *Machine Learning*. 2001;45(1): 5–32. <https://doi.org/10.1023/A:1010933404324>.
48. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research: JMLR*. 2019;20: 177.

A Supplementary Details

A.1 Isolation Forest Mathematical Foundation

The Isolation Forest algorithm constructs an ensemble of isolation trees (iTrees) where each tree is built by recursively partitioning the data. For a given data point $\mathbf{x}$, the anomaly score is calculated as:

$$s(\mathbf{x}, n) = 2^{-\frac{E(h(\mathbf{x}))}{c(n)}}$$

where:

- $E(h(\mathbf{x}))$ is the average path length of observation $\mathbf{x}$ over all isolation trees
- $c(n)$ is the average path length of unsuccessful search in a Binary Search Tree (BST) given n points
- $c(n) = 2H(n-1) - \frac{2(n-1)}{n}$, where H is the harmonic number

The average path length for the unsuccessful search in BST provides normalisation and comparison baseline. When $E(h(\mathbf{x}))$ is close to $c(n)$, the instance is likely normal; when $E(h(\mathbf{x}))$ is significantly smaller than $c(n)$, the instance is likely anomalous.

A.2 Feature Importance Analysis

Dual-method feature ranking identified Zga (GA z-score within a pregnancy) as the top predictive feature by permutation importance (47,48) (see Table S1), while effect size analysis highlighted Ozscore (Z-score for gestation age across pregnancies per outcome) and EGA Max (Maximum expected gestation per outcome) as having the largest statistical differences in adverse pregnancy outcomes (See Table S2). Refer to the feature description table in the methods section 2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementarytablesS1andS2.pdf](#)