Diagnostic Accuracy of a Novel Machine Learning Algorithm to Estimate Gestational Age

Analysis Plan
Version 1.1
08 June 2022

Funded by:
The Bill and Melinda Gates Foundation
OPP 1191684

Principal Investigators:
Joan T. Price, MD, MPH
Bellington Vwalika, MBChB, MMed, MSc

Study Registration:
ClinicalTrials.gov
NCT05433519

For each pregnancy, the gestational age (GA) estimated at the first ultrasound establishes the estimated due date (EDD) and is used to define the "ground truth" gestational age for all visits that follow. Then, during study visits 1, 2, and 3, we will assess the diagnostic accuracy of both the index test (i.e., the FAMLI AI algorithm interpreting blind sweep data) and the clinical reference standard (i.e., standard fetal biometry performed by a trained sonographer). Accuracy will be estimated for both the index test and the clinical standard at each visit by comparing the test's gestational age estimate to the previously determined ground truth gestational age.

The Diagnostic Accuracy (DXA) Study primary outcome is the mean difference in absolute errors in gestational age assessment for the index test versus the clinical reference standard (expert biometry) conducted in the primary evaluation window (visit 1), defined in the protocol as between 14 $^{0}/_{7}$ and 27 $^{6}/_{7}$ gestational weeks, inclusive. The absolute errors, in days, for the index test and expert biometry are measured with respect to the gold standard assessment taken prior to 14 weeks of gestation. We will estimate a Wald-type 95% confidence interval for this difference. Specifically, let m be the mean of the difference in absolute errors for each participant (i.e., $m = n^{-1} \sum_{i=1}^{n} |y_{1(i)}| - |y_{2(i)}|$, where $y_{j(i)}$ is the error, in days from gold standard, for method j and participant i), and σ be the standard error of m, then the 95% confidence interval is given by $m \pm 1.96\sigma$. A negative or positive mean difference whose 95% confidence interval does not include zero indicates the index test is superior or inferior to expert biometry, respectively. A difference whose 95% confidence interval is contained in the prespecified equivalence range of -2 to 2 days indicates the index test is equivalent to expert biometry. To establish equivalence, two one-sided statistical tests on the difference between the mean absolute error of the index test (μ_t) and the mean absolute error the clinical reference standard (μ_r) will be carried out based on a predefined margin c = 2 days i.e. H_{01}: $\mu_t - \mu_r \leq -c$ and H_{02}: $\mu_t - \mu_r \geq c$.[1]

To ensure fair comparison, the analysis will be restricted to only those cases where both the index test and clinical reference standard (expert biometry) are available; hence, we have paired estimates of the two methods. This means that for the primary study outcome, the difference in the MAEs of the two approaches corresponds to the mean of the pairwise difference in absolute errors, m. Study visits that are missing a gestational age estimate from either the index test or the clinical reference standard will be excluded. Potential reasons for missingness include when the index test model indicates a failure to calculate GA based on the blind sweeps of that study, patient non-attendance of a scheduled study visit, and data loss. By design, participants who experience a pregnancy loss after screening but before visit 1 will be excluded from the DXA Study (and therefore from analysis) and replaced with other participants to be enrolled.

As secondary analysis, we will present the difference in root mean squared error and Wald-type 95% confidence interval. Also, as secondary analysis, we will plot the empirical cumulative distribution function for the absolute error produced by the index test and expert biometry. We will present the difference in proportions with absolute error below 7 and 14 days between the index test and expert biometry, along with Wald-type 95% confidence intervals. The primary analysis compares the index test to the clinical reference standard in the pre-defined primary evaluation window (Figure 1). In secondary analyses we will make the same comparison in two secondary windows, defined (a) between $28\ {}^{0}/_{7}$ and $36\ {}^{6}/_{7}$ weeks of gestation, inclusive (visit 2) and (b) between $37\ {}^{0}/_{7}$ weeks and $40\ {}^{6}/_{7}$ weeks, inclusive (visit 3).

Figure 1: Gestational age definitions of the primary and secondary evaluation windows

					Enroll									Primary Evaluation Window														Secondary Evaluation Windows												
GA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
					$<12^0$									14^0 to 27^6														28^0 to 36^6									37^0 to 40^6			

While the primary analysis will compare the index test interpreting blind sweep data as generated by an untrained (novice) sonographer to expert biometry overall, we will also compare the model performance when the blind ultrasound sweeps are generated by an expert sonographer. As in our primary analyses, these analyses will assess model error and clinical reference standard error by comparing estimates to previously established ground truth. Additional secondary analyses will include comparisons in prespecified subsets defined by geography (i.e., Zambia vs North Carolina), medical and obstetrical risk factors (maternal obesity, hypertension, diabetes, fetal growth restriction, oligohydramnios, and polyhydramnios).

Because both the index and reference standard tests are conducted on all women, comparisons of approaches are naturally controlled for all possible time-fixed confounding factors. To account for differential loss to follow-up, if loss to follow-up exceeds 9% and selection bias is suspected, we will employ inverse probability of censoring weights. Censoring weights will be estimated using a pooled logistic regression model fit by maximum likelihood.[2] We will specify *a priori* a set of possible common causes including demographics, medical, and obstetrical factors, for inclusion as covariates in the model. A list of possible covariates is shown below in Table 1. Continuous covariates will be included using restricted quadratic splines.[3]

Table 1: Possible covariates for inverse probability of censoring weights modeling

Age, years	BMI, kg/m2	Estimated gestational age at enrollment, weeks
Education, years	HIV serostatus*	
Marital/partner status	Syphilis serostatus	Transvaginal cervical length, cm
Gravidity (total pregnancies)	Hemoglobin, mg/dL	Study site
Parity (prior deliveries)	Bacteriuria	
Prior miscarriage	Alcohol use in pregnancy	
Prior stillbirth	Tobacco use in pregnancy	
Prior preterm birth	Illicit drug use in pregnancy	

*We will also consider inclusion of HIV-specific covariates through interaction terms (e.g., timing of HIV diagnosis and antiretroviral therapy initiation, antiretroviral therapy regimen, viral load, CD4 count)

References

1. Blair RC, Cole SR. Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods.* 2002;1(1):18.

2. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Statistics in medicine*. 2009 May 30;28(12):1725-38.

3. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ, Jr. Splines for trend analysis and continuous confounder control. *Epidemiology.* 2011;22(6):874-875.