JAMA | Original Investigation | AI IN MEDICINE

Diagnostic Accuracy of an Integrated AI Tool to Estimate Gestational Age From Blind Ultrasound Sweeps

Jeffrey S. A. Stringer, MD; Teeranan Pokaprakarn, PhD; Juan C. Prieto, PhD; Bellington Vwalika, MD, MSc; Srihari V. Chari, MPH; Ntazana Sindano, BSc; Bethany L. Freeman, MSW, MPH; Bridget Sikapande, MSc; Nicole M. Davis, MPH; Yuri V. Sebastião, PHD; Nelly M. Mandona, MBChB; Elizabeth M. Stringer, MD, MSc; Chiraz Benabdelkader, PhD; Mutinta Mungole, BSc; Filson M. Kapilya, BSc; Nariman Almnini, MD; Arieska N. Diaz, BS; Brittany A. Fecteau, AAS; Michael R. Kosorok, PhD; Stephen R. Cole, PhD; Margaret P. Kasaro, MBChB, MPH

IMPORTANCE Accurate assessment of gestational age (GA) is essential to good pregnancy care but often requires ultrasonography, which may not be available in low-resource settings. This study developed a deep learning artificial intelligence (AI) model to estimate GA from blind ultrasonography sweeps and incorporated it into the software of a low-cost, battery-powered device.

OBJECTIVE To evaluate GA estimation accuracy of an AI-enabled ultrasonography tool when used by novice users with no prior training in sonography.

DESIGN, SETTING, AND PARTICIPANTS This prospective diagnostic accuracy study enrolled 400 individuals with viable, single, nonanomalous, first-trimester pregnancies in Lusaka, Zambia, and Chapel Hill, North Carolina. Credentialed sonographers established the "ground truth" GA via transvaginal crown-rump length measurement. At random follow-up visits throughout gestation, including a primary evaluation window from 14 0/7 weeks' to 27 6/7 weeks' gestation, novice users obtained blind sweeps of the maternal abdomen using the AI-enabled device (index test) and credentialed sonographers performed fetal biometry with a high-specification machine (study standard).

MAIN OUTCOMES AND MEASURES The primary outcome was the mean absolute error (MAE) of the index test and study standard, which was calculated by comparing each method's estimate to the previously established GA and considered equivalent if the difference fell within a prespecified margin of ±2 days.

RESULTS In the primary evaluation window, the AI-enabled device met criteria for equivalence to the study standard, with an MAE (SE) of 3.2 (0.1) days vs 3.0 (0.1) days (difference, 0.2 days [95% CI, −0.1 to 0.5]). Additionally, the percentage of assessments within 7 days of the ground truth GA was comparable (90.7% for the index test vs 92.5% for the study standard). Performance was consistent in prespecified subgroups, including the Zambia and North Carolina cohorts and those with high body mass index.

CONCLUSIONS AND RELEVANCE Between 14 and 27 weeks' gestation, novice users with no prior training in ultrasonography estimated GA as accurately with the low-cost, point-of-care AI tool as credentialed sonographers performing standard biometry on high-specification machines. These findings have immediate implications for obstetrical care in low-resource settings, advancing the World Health Organization goal of ultrasonography estimation of GA for all pregnant people.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT05433519

Editorial page 626

Multimedia

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Jeffrey S. A. Stringer, MD, UNC School of Medicine, UNC Global Women's Health, 327 Health Sciences Library, 335 S Columbia St, Chapel Hill, NC 27599-7577 (jeffrey_stringer@med.unc.edu).

JAMA. 2024;332(8):649-657. doi:10.1001/jama.2024.10770
Published online August 1, 2024.

Obstetrical sonography is a cornerstone of modern pregnancy care.[1] Among its many capabilities is the ability to obtain accurate measurements of fetal structures, which in turn are used to estimate gestational age (GA).[2,3] Obstetrical clinicians use GA to guide various aspects of antenatal care, such as when to screen for gestational diabetes[4] and when to administer certain vaccines to maximize maternal and neonatal benefit.[5] GA also critically informs clinical decision-making, such as whether to provide corticosteroids[6] or neuroprotective magnesium sulfate[7] for anticipated preterm delivery and whether clinician-initiated delivery is appropriate for a given condition.[8,9]

The World Health Organization recommends that all pregnant people receive at least 1 ultrasonography examination prior to 24 weeks.[10] Although this policy recommendation remains largely aspirational in many low- and middle-income countries (LMICs), recent advances in ultrasonography hardware[11,12] and artificial intelligence (AI)–enabled medical image analysis[13,14] could facilitate broader access to this critical diagnostic tool. In 2022, a deep learning algorithm developed in an international study of 4695 pregnant volunteers that could estimate GA from blindly obtained ultrasound sweeps of the gravid abdomen was examined.[15] Here, in a separate cohort, the diagnostic accuracy of that algorithm is reported when incorporated into the software of a low-cost battery-powered device and used by clinicians with no formal training in sonography.

Key Points

Question Can novice clinicians accurately estimate gestational age using a low-cost, battery-powered ultrasonography probe with integrated artificial intelligence (AI) image interpretation?

Findings This prospective study enrolled 400 pregnant individuals with due dates confirmed by first-trimester ultrasonography. At follow-up visits randomly assigned throughout gestation, novice clinicians using an AI-enabled device estimated gestational age as accurately as credentialed sonographers using traditional ultrasonography devices (difference, 0.2 days).

Meaning Obstetrical care in low-resource settings may benefit from reliable gestational age assessment using AI integration with point-of-care ultrasonography.

Methods

This prospective diagnostic accuracy study enrolled 400 pregnant individuals with viable, single, nonanomalous, first-trimester pregnancies in Lusaka, Zambia and Chapel Hill, North Carolina. This study received approval from the institutional review board at the University of North Carolina, the University of Zambia Biomedical Research Ethics Committee, the Zambia Medicines Regulatory Authority, and the Zambia National Health Research Authority before initiation. An external auditor conducted quarterly site visits in both North Carolina and Zambia to ensure compliance with the study protocol, standard operating procedures, International Conference on Harmonization Good Clinical Practice, and US 45 CFR 46 regulations.

Credentialed sonographers established the "ground truth" GA via transvaginal crown-rump length measurement.[16] Participants were then assigned follow-up visits at random dates within a primary GA evaluation window (14 0/7 to 27 6/7 weeks' gestation) and 2 secondary windows to ensure observations were evenly spaced in an unbiased manner throughout the pregnancy. At each follow-up visit, novice users with no prior training in sonography assessed GA with blind sweeps of the maternal abdomen using the AI-enabled device (index test). The technology for the index test comprised a previously described deep learning model[15] incorporated into the software of the Butterfly IQ+ handheld ultrasonography device (Butterfly Networks, Inc). To facilitate integration into the Butterfly IQ+ software, we made modifications to optimize the model for real-time inference on a mobile device. We also incorporated a fail-safe mechanism that required the user to repeat collection of blind sweeps that did not reach a certain quality threshold (see Supplement 1).

The study employed obstetrics-trained sonographers, each credentialed by the operant authority in their country (the American Registry for Diagnostic Medical Sonography or the Health Professions Council of Zambia). The credentialed sonographer used a high-specification ultrasound machine to assess GA with fetal biometry (study standard). The index test was performed first, using a software version that did not display the calculated GA at the completion of the procedures. Index test users were not allowed to consult with study sonographers while using the tool. During study implementation, both novice users and credentialed study sonographers were blinded to the participant's ground truth GA.

The study was conducted at the University Teaching Hospital and the Kamwala District Health Centre in Lusaka, Zambia and at the University of North Carolina Vilcom Center Clinic in Chapel Hill, North Carolina. We included people who (1) were 18 years or older, (2) had a viable intrauterine pregnancy at less than 14 0/7 weeks' gestation, (3) provided written informed consent, (4) intended to remain in the current geographical area of residence for the duration of study, and (5) were willing to adhere to study procedures. We excluded people who (1) had a body mass index (BMI) greater than or equal to 40, (2) were pregnant with twins or higher-order multiples, (3) had a known major fetal anomaly, or (4) had any social or medical condition that would make study participation unsafe or complicate data interpretation.

With anticipation that the principal use of the index test would be in LMIC settings in which initial presentation for pregnancy care typically occurs later in gestation[17] than in North America and Europe, we defined a primary evaluation window from 14 0/7 to 27 6/7 gestational weeks. This window corresponds to a range that would capture 85% of individuals attending their first antenatal visit in LMICs.[18] For secondary analyses, we defined a secondary evaluation window (28 0/7 to 36 6/7 gestational weeks) and a tertiary evaluation window (37 0/7 to 40 6/7 gestational weeks).

The study employed randomization to assign a participant's visit schedule and thus the GA at assessment within each evaluation window. A statistician not involved in study implementation designed the randomization scheme and pregenerated each participant's visit schedule prior to study commencement. The scheme did not allow a participant who was assigned to the last week in an evaluation window to also be assigned to the first week in the subsequent window (ie, to have 2 study visits only 1 week apart); this was the only constraint on the randomization.

Index Test

The index test was designed for use by novice clinicians without prior training in sonography. Before the study commenced, novice users were identified at each site (eTable 7 in Supplement 2) and underwent a 1-day training session. The curriculum covered software navigation, patient positioning, gel application, probe orientation and pressure, and blind sweep collection. Half of the training day was spent getting hands-on experience with patients in the research clinic using the tool under the supervision of an experienced sonographer.

The index test began with the novice user assessing the symphysis-fundal height and entering the resultant measurement (in centimeters) into the device software. This allowed the tool to set the number of required sweeps and configure the ultrasound probe's depth and gain settings. The software then guided the user through collection of a series of 10-second blind sweep videos (eFigure 1 in the Supplement 2). Although the software offered an instructional animation demonstrating probe movement, it did not display real-time ultrasonography images (see **Video**).

Study Standard

The study standard for GA assessment was fetal biometry.[16] At each study visit, a credentialed sonographer obtained 2 separate measurements of the fetal head circumference, biparietal diameter, abdominal circumference, and femur length on a high-specification ultrasonography machine (General Electric Healthcare). The mean of the 2 measures was used to calculate the GA on that day using either the 2-parameter Intergrowth 21 formula[3] (Zambia) or 4-parameter Hadlock formula[2] (North Carolina). Consistent with previous publications[15,19] we explored the impact of different biometry formulas on outcomes through sensitivity analyses.

Study Outcomes

This study assessed estimation error of the index test and study standard by comparing each test's estimate with the ground truth GA previously established in early pregnancy. Our primary outcome measure was the difference in mean absolute error (MAE) between the index test and the study standard, assessed in the primary evaluation window. Secondary outcome measures were the difference in MAE between the 2 tests assessed in the secondary and tertiary evaluation windows, the difference in root mean square error assessed in all 3 windows, and the difference in the proportion of studies correctly classified within 7 and 14 days of ground truth in all 3 windows.

Statistical Approach

We hypothesized that the index test would be equivalent to the study standard and, through consultation with experts in North Carolina and Zambia, established a mean estimation error no worse or better than 2 days as the equivalency margin[20] for this study. We used Monte Carlo simulation to establish a sample size that yielded at least 95% power for the ±2-day equivalency margin and type I error of 2.5% (further details are available in Supplement 3).

We calculated a 95% CI for the primary outcome. A difference for which the 2-sided 95% CI is contained entirely within the prespecified range of −2 to 2 days would indicate that the index test is equivalent to the study standard. To establish equivalence, 2 one-sided statistical tests on the difference between the MAE of the index test and the MAE of the study standard were carried out based on the predefined margin. As secondary analyses, we present the difference in root mean square error and its 95% CI. We also plot the empirical cumulative distribution function for the absolute error produced by the index test and expert biometry. We then present the difference in percentages with absolute error below 7 and 14 days between the index test and study standard, along with Wald-type 95% CIs.

Subgroup analyses prespecified in our statistical analysis plan included geographic location and high BMI (≥30). Additionally, because many LMICs do not have ultrasound biometry widely available, we conducted an exploratory analysis comparing the performance of the index test with that of the de facto study standard in these settings: patient-reported last menstrual period[21] and measurement of the symphysis-fundal height.[22]

Results

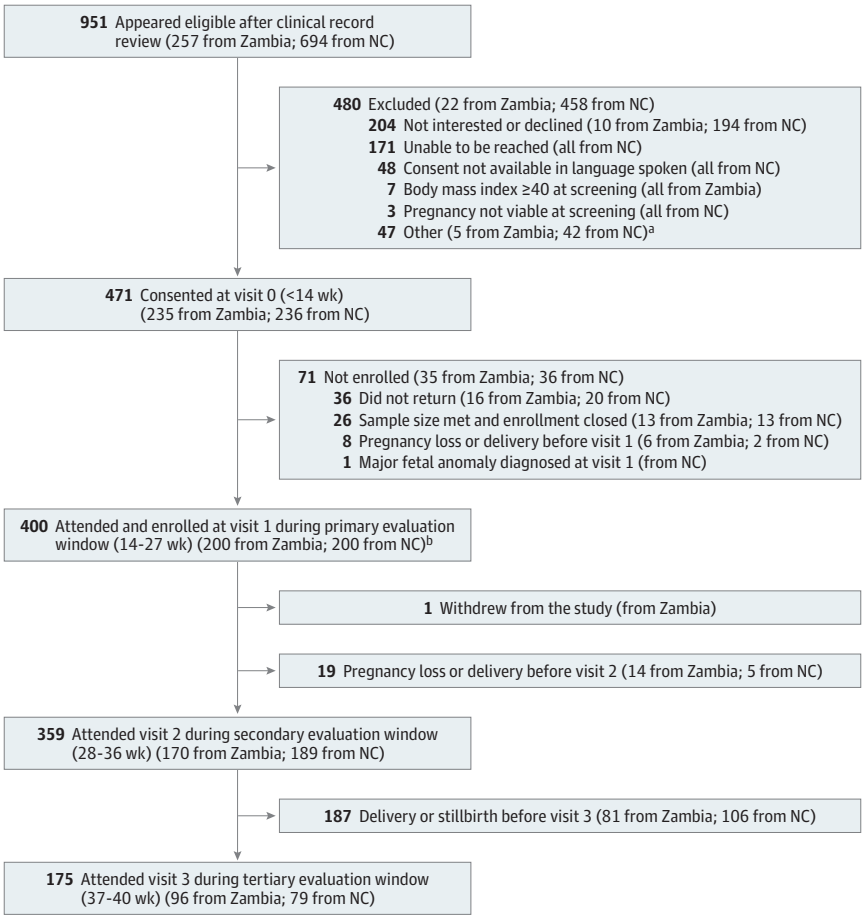
Between July 27, 2022, and April 10, 2023, a total of 951 individuals who appeared eligible to participate were identified through clinical record review. Of these individuals, 480 were excluded because they were either unable to be reached, found to be ineligible after further investigation, or not interested in participation (**Figure 1**). The remaining 471 provided informed consent to participate and were randomly assigned dates for follow-up visits in the 3 evaluation windows. On May 31, 2023, the 400th participant attended visit 1 (primary evaluation window) and study enrollment was closed.

The 400 study participants had a median (IQR) age of 29 (25-33) years, a median (IQR) of 13 (10-16) years of education, and a median (IQR) BMI of 25.9 (22.6-29.9). Overall, 252 participants (63%) were parous and 25 (all in Zambia; 8.0%) were HIV-seropositive. Compared with the North Carolina cohort, participants in Zambia were younger with lower BMI, lower rates of chronic hypertension and diabetes, and higher parity (**Table 1**). No adverse events were attributed to the index test or the reference standard at any visit.

Primary Evaluation Window (14-27 Gestational Weeks)

All 400 participants were assessed with both the index test and study standard during the primary evaluation window. In 1 case

Figure 1. Participant Flow



NC indicates North Carolina.

[a]Most candidates in the "other" category were excluded because of unreliable transportation and concern by the study team that they may not be able to attend all scheduled visits. There were 2 individuals excluded for clinical reasons in North Carolina. One person had a colostomy that might interfere with the novice sweep procedure and another had large uterine myomas.

[b]The index test did not produce a gestational age (GA) estimate for 1 participant during the primary evaluation window, thus diagnostic accuracy is calculated among 399 individuals. There were no instances of failure to calculate in the secondary and tertiary windows. The reference standard produced a result for all evaluated patients in all 3 windows.

(0.25%) the index test failed to produce a GA estimate, while standard fetal biometry was successfully obtained from all 400 participants. Among the 399 individuals from whom paired assessments are available, the index test MAE (SE) was 3.19 (0.13) days, compared with 3.03 (0.12) days for the study standard (difference, 0.16 [95% CI, −0.14 to 0.45] days; **Table 2**; **Figure 2**), meeting the predefined equivalency margin. The proportion of assessments correctly classified within 7 days of the ground truth GA were comparable between the 2 methods (90.7% for the index test vs 92.5% for the study standard; difference, −1.8% [95% CI, −5.0% to 1.5%]). Both tests were highly accurate for GA estimates within a 14-day range, each misestimating 1 (distinct) participant by more than 14 days (99.8% for the index test vs 99.8% for the study standard; difference, 0% [95% CI, 0% to 0.7%]).

Table 2 displays the index test performance by geography and BMI, without adjustment to account for multiple comparisons. There was a similar difference in MAE between the index and study standard by site (Zambia [n = 199]: −0.18 [95% CI, −0.56 to 0.20] days; North Carolina [n = 200]: 0.49 [95% CI, 0.05 to 0.94] days). Among the subgroup whose first-visit BMI was greater than or equal to 30 (n = 97), the difference between tests was 0.70 (95% CI, 0.07 to 1.33) days (Table 2).

eTables 5 and 6 in Supplement 2 display results from 2 planned sensitivity analyses that assessed all biometry using a single uniform formula (ie, one analysis applying Intergrowth 21 to both countries and a second applying Hadlock to both). These analyses revealed that employing distinct formulas by site did not materially influence findings or conclusions.

Secondary (28-36 Gestational Weeks) and Tertiary (37-40 Gestational Weeks) Evaluation Windows

The secondary evaluation window spanned the 9-week interval from 28 0/6 to 36 6/7 weeks' gestation. Between scheduled visits in the primary and secondary evaluation windows, 1 participant formally withdrew from the study and 19 had a miscarriage or preterm delivery, reducing the expected attendance in the secondary window to 380 participants (Figure 1). Of these participants, 359 (94.5%) attended as anticipated. In all 359 participants, both the index test and clinical standard produced a GA estimate. During the secondary window, the index test MAE (SE) was 6.07 (0.26) days, compared with 7.12 (0.30) days for the study standard (difference, −1.06 [95% CI, −1.72 to −0.40] days; eTable 1 and eFigure 2 in Supplement 2), meeting the study definition of

Table 1. Study Participant Characteristics at Enrollment

	Overall (N = 400)[a]	Zambia (n = 200)[a]	North Carolina (n = 200)[a]
Demographic characteristics			
Maternal age, median (IQR), y	29.0 (25.0-33.0)	27.0 (23.0-31.0)	32.0 (27.5-34.0)
Education, median (IQR), y	13.0 (10.0-16.0)	10.0 (9.0-12.0)	16.0 (14.0-18.0)
Living with partner	347 (86.8)	165 (82.5)	182 (91.0)
Obstetrical history			
Gestational age at first ultrasound, median (IQR), wk	11.7 (10.0-12.9)	10.7 (8.9-12.6)	12.4 (11.1-13.3)
Parous	252 (63.0)	137 (68.5)	115 (57.5)
Prior miscarriage	104/252 (41.3)	44/137 (32.1)	60/115 (52.2)
Prior preterm birth	51/252 (20.2)	26/137 (19.0)	25/115 (21.7)
Hypertension during previous pregnancy	23 (5.8)	6 (3.0)	17 (8.5)
Preeclampsia/eclampsia during previous pregnancy	14 (3.5)	1 (0.5)	13 (6.5)
Maternal health at enrollment			
BMI, median (IQR)	25.9 (22.6-29.9)	24.9 (22.0-29.0)	27.0 (23.1-31.4)
Mid-upper arm circumference, median (IQR), cm	28.0 (25.5-30.5)	28.0 (25.5-31.0)	27.8 (25.5-30.0)
No.	390	199	191
Hypertension outside of pregnancy	22 (5.5)	4 (2.0)	18 (9.0)
Currently taking BP medications	11/22 (50.0)	3/4 (75.0)	8/18 (44.4)
Diabetes outside of pregnancy	13 (3.3)	1 (0.5)	12 (6.0)
Currently taking oral or injectable diabetes medications	11/13 (84.6)	1/1 (100.0)	10/12 (83.3)
Hemoglobin, median (IQR), mg/dL	13.0 (12.2-13.5)	13.0 (11.6-13.8)	13.0 (12.2-13.5)
No.	164	19	145
HIV seropositive	25/311 (8.0)	25/183 (13.7)	0/128
Syphilis seropositive	9/293 (3.1)	9/167 (5.4)	0/126
Abnormal urine dipstick test[b]	16/135 (11.9)	7/62 (11.3)	9/73 (12.3)
Alcohol use	13 (3.3)	10 (5.0)	3 (1.5)
Tobacco use	9 (2.3)	1 (0.5)	8 (4.0)

Abbreviations: BMI, body mass index; BP, blood pressure.

[a] Data are presented as number or number/total number (percentage) of participants unless otherwise indicated.

[b] Abnormal defined as ≥1+ leukocyte esterase or + nitrites.

Table 2. Gestational Age Estimation Assessed in the Primary Evaluation Window (14 0/7 to 27 6/7 Weeks' Gestation)[a]

	Index test (95% CI)	Study standard (95% CI)	Difference (95% CI)
Mean absolute error, d	3.19 (2.93 to 3.45)	3.03 (2.79 to 3.27)	0.16 (−0.14 to 0.45)
Mean error, d	1.29 (0.90 to 1.67)	0.64 (0.27 to 1.02)	
Root mean square error, d	4.14 (3.81 to 4.47)	3.89 (3.57 to 4.22)	0.24 (−0.12 to 0.61)
Absolute error <7 d, %	90.7 (87.9 to 93.6)	92.5 (89.9 to 95.1)	−1.8 (−5.0 to 1.5)
Absolute error <14 d, %	99.8 (99.3 to 100.0)	99.8 (99.3 to 100.0)	0.00 (−0.7 to 0.7)
Zambia site mean absolute error, d (n = 199)	3.02 (2.65 to 3.39)	3.20 (2.83 to 3.57)	−0.18 (−0.56 to 0.20)
North Carolina site mean absolute error, d (n = 200)	3.35 (2.99 to 3.72)	2.86 (2.55 to 3.17)	0.49 (0.05 to 0.94)
Mean absolute error among those with BMI >30, d (n = 97)	3.78 (3.23 to 4.33)	3.07 (2.51 to 3.64)	0.70 (0.07 to 1.33)

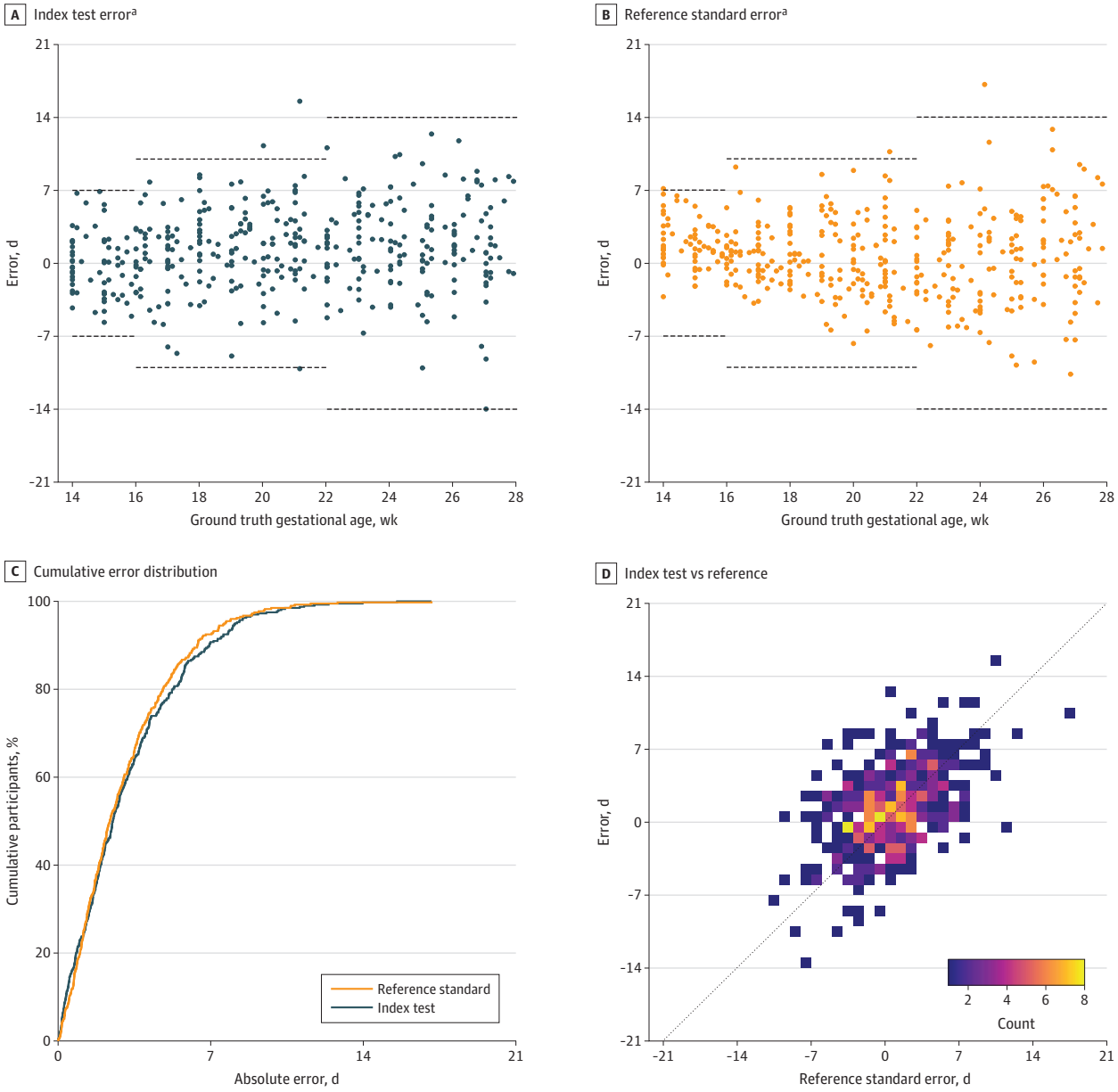
Abbreviation: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared).

[a] The index test did not produce a result in 1 participant; this observation was excluded from the analysis.

equivalence. eTable 2 in Supplement 2 displays the percentage of assessments correctly classified within 7 and 14 days: 64.4% (95% CI, 59.4%-69.3%) and 91.4% (95% CI, 88.5%-94.3%), respectively, for the index test compared with 57.1% (95% CI, 52.0%-62.2%) and 87.2% (95% CI, 83.7%-90.6%), respectively, for the study standard (difference: 7.2% [95% CI, 0.7%-13.8%] for within 7 days and 4.2% [95% CI, 0.1%-8.3%] for within 14 days).

The tertiary evaluation window spanned the 4-week interval from 37 0/6 to 40 6/7 weeks' gestation. Of the 380 participants with a continuing, viable pregnancy at the second visit, 187 either delivered or experienced a stillbirth before their scheduled visit in the tertiary evaluation window. Thus, the expected attendance for the third visit was 193 participants, of whom 175 (91%) attended (Figure 1). In all participants both the index test and study standard produced a GA estimate; however, neither test performed particularly well. The index test had a MAE (SE) of 11.54 (0.49) days, compared with 9.10 (0.54) days for the study standard (difference, 2.43 [95% CI, 1.19-3.68] days; eTable 2 and eFigure 3 in Supplement 2).

Figure 2. Performance of Index Test vs Study Reference Standard in the Primary Evaluation Window



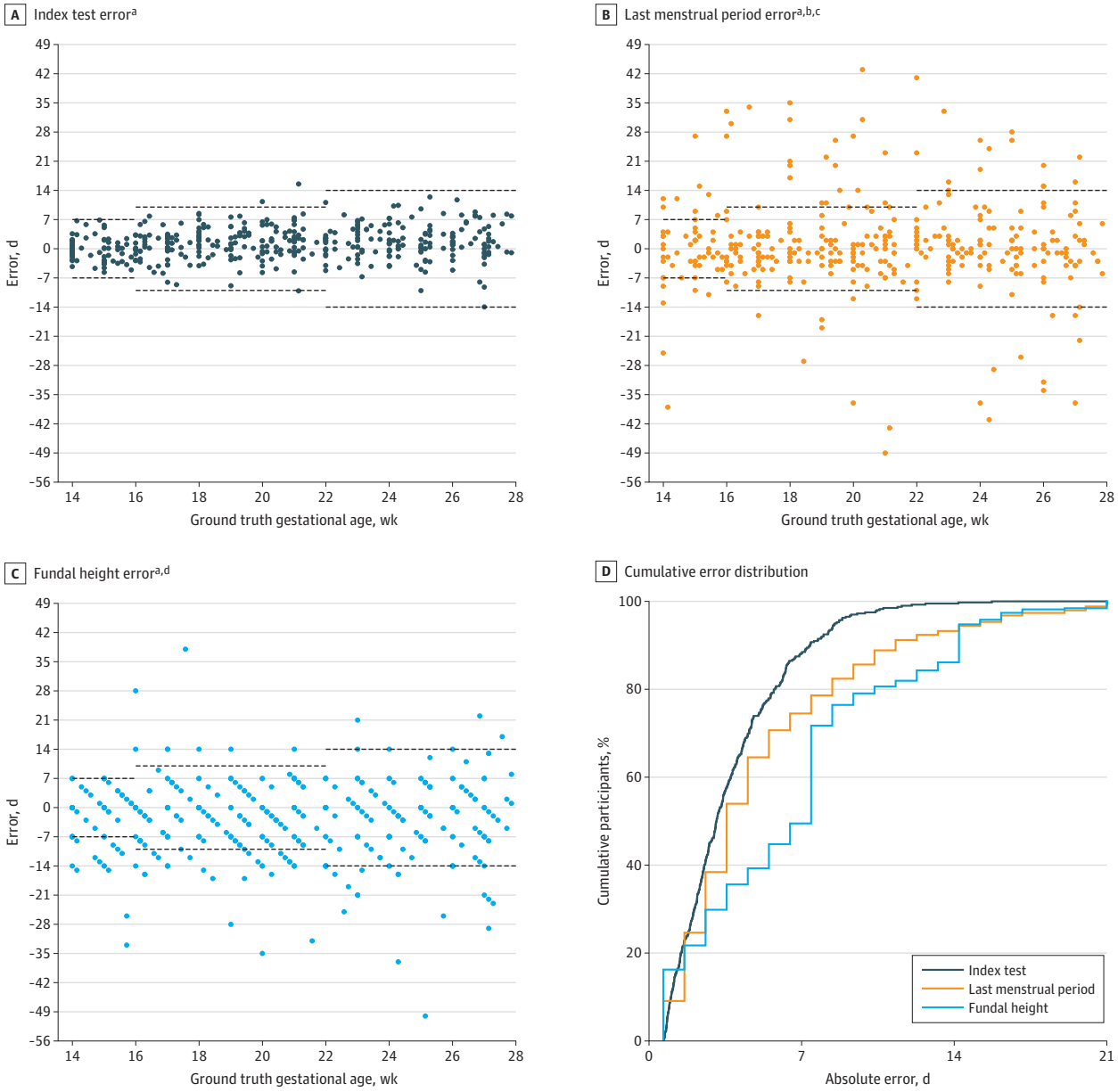
[a]Dashed horizontal lines represent expected error bounds of ultrasound biometry according to the American College of Obstetricians and Gynecologists.[16]

eTable 2 in Supplement 2 shows the percentage of assessments correctly classified within 7 and 14 days: 28.0% and 66.3%, respectively, for the index test and 46.3% and 74.3%, respectively, for the study standard (difference: −18.3% [95% CI, −27.9% to −8.64%] for within 7 days and −8.0% [95% CI, −16.7% to 0.74%] for within 14 days; eTable 2 in Supplement 2).

Index Test vs de Facto GA Assessment Standards

In an exploratory analysis, performance of the index test during the primary evaluation window was compared with the de facto GA assessment standard in many LMIC settings: last menstrual period and symphysis-fundal height. Among the 399 individuals from whom an index test estimate was available during the primary evaluation window, 23 (5.8%) could not recall their last menstrual period and were excluded. Among the remaining 376 participants, the MAE (SE) was 3.20 (0.14) days for the index test compared with 7.44 (0.51) days for last menstrual period (difference, −4.24 [95% CI, −5.27 to −3.20] days; **Figure 3**; eTable 3 in Supplement 2). The symphysis-fundal height could not be assessed because the uterus was not palpable in 4 of the 399 individuals from whom an index test estimate was available. Of the remaining 395 participants, the index test had an MAE (SE) of 3.18 (0.13) days compared with 7.06 (0.34) days for fundal height (difference, −3.88 [95% CI, −4.61 to −3.15] days; Figure 3; eTable 4 in Supplement 2).

Figure 3. Performance of Index Test vs Symphysis-Fundal Height and Last Menstrual Period in the Primary Evaluation Window



[a]Dashed horizontal lines represent expected error bounds of ultrasound biometry according to the American College of Obstetricians and Gynecologists.[16]

[b]Twenty-three participants could not recall their last menstrual period and were excluded.

[c]Three participants (2 with last menstrual period error >49 days and 1 with last menstrual period error ≤49 days) and are not represented on this plot.

[d]Four participants had a nonpalpable uterine fundus and were excluded.

Discussion

This prospective, 2-country diagnostic accuracy study provides evidence that an AI-enabled ultrasonography tool used by novice clinicians with 1 day of training can provide GA estimates that are as accurate as credentialed sonographers performing standard fetal biometry. Specifically, over the critical GA window during which most people in LMIC settings attend their first antenatal visit (14-27 weeks' gestation), the index test met the predefined criteria for statistical equivalence to a credentialed sonographer using a high-specification machine. Although hypothesis testing was not performed in the subgroup analyses, these findings appear to be consistent across geography (Zambia and North Carolina) and among participants with high BMI (in whom ultrasonography can be more difficult to perform). The index test also met criteria for equivalency in a secondary GA window between

28 and 36 weeks' gestation, whereas results were inconclusive after term (≥37 weeks' gestation).

The selection of evaluation windows was informed by a report of more than 100 000 pregnancies in Zambia, which revealed that 85% of first antenatal visits occur by the end of the primary evaluation window and 97% by the end of the secondary window.[17] These figures are remarkably consistent across LMICs and confirmed by a recent comprehensive review.[18] In line with prior work,[15] the deep learning AI model appears to perform particularly well during the secondary evaluation window (28-36 weeks' gestation), a period during which meaningful variations in fetal size, attributable to pathological or constitutional factors, begin to emerge. Conversely, the model appears to underperform fetal biometry after term gestation (37-40 weeks' gestation) and, although these results are statistically inconclusive, it is not recommended to use this antenatal assessment tool to determine GA at term.

Unlike previous reports in which ultrasonography videos were processed and analyzed on a central server,[15,19] this study demonstrates the feasibility of integrating an AI tool into clinical practice. The deep learning model was incorporated directly into the ultrasonography device software, which runs on an Android tablet computer, allowing image processing, feature extraction, and inference to occur in real time on the local device, facilitating immediate clinical decision-making (see Video). This research has the potential to inform expansion of basic obstetrical ultrasonography, bringing previously unavailable diagnostic capacity to settings in which resources are scarce but clinical disease burden is high.

Several methodological strengths support the validity of these findings. The current study enrolled a socioeconomically diverse cohort whose GA was established by first trimester crown-rump length. In an effort to mitigate expected value bias,[23] both expert sonographers and novice users were blinded to participant ground truth GA and to the results of each other's assessments. The study employed a novel use of randomization to ensure unbiased allocation of participant visits across all possible gestational ages.

Limitations

There are important limitations to this research. First, the study enrolled a general obstetrical population and was not designed to assess performance of the AI-enabled tool among patients with high-risk conditions linked to inaccurate GA dating. Assessing performance of the index test in settings of hypertension, diabetes, and class III obesity—also challenging to traditional ultrasound biometry—will be important future work. Second, although the 3 sites in 2 countries provided socioeconomic diversity, inclusion of more geographic locations could improve generalizability of these findings. Third, because the protocol excluded participants with known fetal anomalies, the accuracy of the tool in such cases cannot be determined.

Conclusions

Between 14 and 37 gestational weeks, low-cost AI-enabled ultrasonography allowed novice users with no prior training in ultrasonography to estimate GA as accurately as credentialed sonographers performing standard biometry on high-specification machines. These findings have immediate implications for obstetrical care in low-resource settings, advancing the World Health Organization goal of ultrasonography estimation of GA for all pregnant people.

ARTICLE INFORMATION

Accepted for Publication: May 17, 2024.

Published Online: August 1, 2024. doi:10.1001/jama.2024.10770

Author Affiliations: Department of Obstetrics and Gynecology, University of North Carolina School of Medicine, Chapel Hill (J. S. A. Stringer, Chari, Freeman, Davis, Sebastião, E. M. Stringer, Benabdelkader, Almnini, Diaz, Fecteau); Department of Biostatistics, University of North Carolina Gillings School of Global Public Health, Chapel Hill (Pokaprakarn); Department of Psychiatry, University of North Carolina School of Medicine, Chapel Hill (Prieto); Department of Obstetrics and Gynaecology, University of Zambia School of Medicine, Lusaka, Zambia (Vwalika, Kasaro); UNC Global Project-Zambia, Lusaka, Zambia (Sindano, Sikapande, Mandona, Mungole, Kapilya); Department of Biostatistics, University of North Carolina Gillings School of Global Public Health, Chapel Hill (Kosorok); Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill (Cole).

Author Contributions: Drs J. Stringer and Kasaro had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: J. Stringer, Pokaprakarn, Freeman, Sebastiao, Kosorok, Cole.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: J. Stringer, Pokaprakarn, Chari, Davis.
Critical review of the manuscript for important intellectual content: Pokaprakarn, Prieto, Vwalika, Chari, Sindano, Freeman, Sikapande, Davis, Sebastiao, Mandona, E. Stringer, Benabdelkader, Mungole, Kapilya, Almnini, Diaz, Fecteau, Kosorok, Cole, Kasaro.
Statistical analysis: J. Stringer, Pokaprakarn, Davis, Sebastiao, Kosorok, Cole.
Obtained funding: J. Stringer.
Administrative, technical, or material support: J. Stringer, Prieto, Vwalika, Chari, Sindano, Freeman, Sikapande, Davis, Mandona, E. Stringer, Mungole, Kapilya, Almnini, Diaz, Fecteau, Kasaro.
Supervision: J. Stringer, Chari, Freeman, Sikapande, Sebastiao, Mandona, Mungole, Kapilya, Kasaro.

Conflict of Interest Disclosures: None reported.

Funding/Support: This work was funded by the Bill and Melinda Gates Foundation (INV003266). Butterfly Systems, Inc donated ultrasonography probes and worked with the investigators to integrate an encrypted version of the deep learning model into their native device software. The machine learning model evaluated in this research was funded by the Bill and Melinda Gates Foundation and is subject to their "Global Access" requirements. As such, the University of North Carolina will make the machine learning model available at no cost in low and middle-income countries (patent No. WO/2023/122326).

Role of the Funder/Sponsor: The Bill and Melinda Gates Foundation and Butterfly Systems, Inc had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The conclusions and opinions expressed in this article are solely those of the authors.

Data Sharing Statement: See Supplement 4.

Additional Contributions: We thank our research participants in Lusaka and Chapel Hill for providing data to this study. We thank Joan T. Price, MD (Rutland Women's Health Care), for her invaluable contributions to the planning and implementation of this study.

REFERENCES

1. Committee on Practice Bulletins—Obstetrics and the American Institute of Ultrasound in Medicine. Practice Bulletin No. 175: ultrasound in pregnancy.

Obstet Gynecol. 2016;128(6):e241-e256. doi:10.1097/AOG.0000000000001815

2. Hadlock FP, Deter RL, Harrist RB, Park SK. Estimating fetal age: computer-assisted analysis of multiple fetal growth parameters. *Radiology*. 1984;152(2):497-501. doi:10.1148/radiology.152.2.6739822

3. Papageorghiou AT, Kemp B, Stones W, et al; International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound Obstet Gynecol*. 2016;48(6):719-726. doi:10.1002/uog.15894

4. World Health Organization. WHO recommendations on antenatal care for a positive pregnancy experience. 2016.Accessed March 15, 2018. https://apps.who.int/iris/bitstream/10665/250796/1/9789241549912-eng.pdf?ua=1

5. American College of Obstetricians and Gynecologists. ACOG Committee Opinion No. 741: maternal immunization. *Obstet Gynecol*. 2018;131(6):e214-e217. doi:10.1097/AOG.0000000000002662

6. Vogel JP, Ramson J, Darmstadt GL, et al. Updated WHO recommendations on antenatal corticosteroids and tocolytic therapy for improving preterm birth outcomes. *Lancet Glob Health*. 2022;10(12):e1707-e1708. doi:10.1016/S2214-109X(22)00434-X

7. American College of Obstetricians and Gynecologists. Committee Opinion No. 455: magnesium sulfate before anticipated preterm birth for neuroprotection. *Obstet Gynecol*. 2010;115(3):669-671. doi:10.1097/AOG.0b013e3181d4ffa5

8. American College of Obstetricians and Gynecologists. Gestational hypertension and preeclampsia: ACOG practice bulletin, number 222. *Obstet Gynecol*. 2020;135(6):e237-e260. doi:10.1097/AOG.0000000000003891

9. American College of Obstetricians and Gynecologists. Practice bulletin no. 146: management of late-term and postterm pregnancies. *Obstet Gynecol*. 2014;124(2 Pt 1):390-396. doi:10.1097/01.AOG.0000452744.06088.48

10. World Health Organization. Maternal and fetal assessment update: imaging ultrasound before 24 weeks of pregnancy. March 28, 2022. Accessed June 6, 2024. https://www.who.int/publications/i/item/9789240046009

11. Venkatayogi N, Gupta M, Gupta A, et al. From seeing to knowing with artificial intelligence: a scoping review of point-of-care ultrasound in low-resource settings. *Appl Sci (Basel)*. 2023;13(8427). doi:10.3390/app13148427

12. Ranger BJ, Bradburn E, Chen Q, Kim M, Noble JA, Papageorghiou AT. Portable ultrasound devices for obstetric care in resource-constrained environments: mapping the landscape. *Gates Open Res*. Published online December 6, 2023. doi:10.12688/gatesopenres.15088.1

13. Jost E, Kosian P, Jimenez Cruz J, et al. Evolving the era of 5D ultrasound? a systematic literature review on the applications for artificial intelligence ultrasound imaging in obstetrics and gynecology. *J Clin Med*. 2023;12(21):6833. doi:10.3390/jcm12216833

14. Chen X, Wang X, Zhang K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*. 2022;79:102444. doi:10.1016/j.media.2022.102444

15. Pokaprakarn T, Prieto JC, Price JT, et al. AI estimation of gestational age from blind ultrasound sweeps in low-resource settings. *NEJM Evid*. 2022;1(5). doi:10.1056/EVIDoa2100058

16. American College of Obstetricians and Gynecologists. Committee Opinion No 700: methods for estimating the due date. *Obstet Gynecol*. 2017;129(5):e150-e154. doi:10.1097/AOG.0000000000002046

17. Chi BH, Vwalika B, Killam WP, et al. Implementation of the Zambia electronic perinatal record system for comprehensive prenatal and delivery care. *Int J Gynaecol Obstet*. 2011;113(2):131-136. doi:10.1016/j.ijgo.2010.11.013

18. Jiwani SS, Amouzou-Aguirre A, Carvajal L, et al. Timing and number of antenatal care contacts in low and middle-income countries: analysis in the Countdown to 2030 priority countries. *J Glob Health*. 2020;10(1):010502. doi:10.7189/jogh.10.010502

19. Viswanathan AV, Pokaprakarn T, Kasaro MP, et al. Deep learning to estimate gestational age from fly-to cineloop videos: a novel approach to ultrasound quality control. *Int J Obstet Gynecol*. 2024;165(3):1013-1021. doi:10.1002/ijgo.15321

20. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG, Group C; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308(24):2594-2604. doi:10.1001/jama.2012.87802

21. Price JT, Winston J, Vwalika B, et al. Quantifying bias between reported last menstrual period and ultrasonography estimates of gestational age in Lusaka, Zambia. *Int J Gynaecol Obstet*. 2019;144(1):9-15. doi:10.1002/ijgo.12686

22. Papageorghiou AT, Ohuma EO, Gravett MG, et al; International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). International standards for symphysis-fundal height based on serial measurements from the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: prospective cohort study in eight countries. *BMJ*. 2016;355:i5662. doi:10.1136/bmj.i5662

23. Drukker L, Droste R, Chatelain P, Noble JA, Papageorghiou AT. Expected-value bias in routine third-trimester growth scans. *Ultrasound Obstet Gynecol*. 2020;55(3):375-382. doi:10.1002/uog.21929