Evaluating the Sensitivity to Change of AI-Feedback in Ultrasound Biometry: A Stratified Randomized Controlled Trial across the Expertise Gradient

Authors & Affiliation: [1, 2, 3]Mary Le Ngo, [4]Chun Kit Wong, [1]Kathinka Marie Hoff Nyborg, Marie Juhl [4]Aasa Feragen, Anders Nymark Christensen [2, 3, 6]Martin Grønnebæk Tolsgaard

[1]Department of Gynecology and Obstetrics, Slagelse Hospital, Slagelse, Denmark

[2] Copenhagen Academy for Medical Education and Simulation (CAMES), Rigshospitalet, Denmark

[3]University of Copenhagen, Faculty of Health Sciences, Copenhagen, Denmark

[4]The Technical University of Denmark, Lyngby, Denmark

[5]Department of Computer Science, Copenhagen University, Copenhagen Denmark

[6]Department of Fetal Medicine, Copenhagen University Hospital, Copenhagen Denmark

Pioneer Center

Trial registrations:

Danish Regional Scientific Ethics Committee: F-250551591

Regional Data Protection Authority: P-2023-15074

Clinical trial:

Protocol number: 1.1 (25-03-2026)

1. Introduction:

In obstetric ultrasound, inter-observer variability remains a significant source of diagnostic uncertainty. While Artificial Intelligence (AI) is frequently proposed as a solution to standardize performance, research often treats clinicians as a homogenous group. However, the clinical utility of AI likely depends on the user's baseline proficiency. To evaluate the true impact of this technology, we must investigate how its implementation interacts with varying levels of professional expertise.

This research is grounded in the Expertise Reversal Effect[1, 2], which suggests that instructional scaffolds—such as real-time AI feedback—that benefit novices may paradoxically impair the performance of experts by disrupting their established, automated workflows. Consequently, we hypothesize the existence of a Goldilocks Zone at the intermediate level of expertise. In this stage, clinicians possess the motor skills to manipulate the probe but have not yet fully internalized the complex schemas of an expert, potentially making them the most receptive to AI-driven corrections.

To identify this threshold, the study utilizes Sensitivity to Change as the primary analytical lens. Rather than focusing solely on static accuracy, we measure the interaction between the AI and the operator to determine the degree to which different clinical users adapt their behavior in response to automated feedback.

2. Study Design:

This study is a prospective, two-arm, Randomized Controlled Trial (RCT) employing a 3x2 design to evaluate the impact of AI-mediated feedback on fetal biometry. 75 participants, stratified into three cohorts based on lifetime scan volume (Novice, Intermediate, and Expert), will be randomized 1:1 to either the AI-Intervention arm or the Standard Care (Control) arm. Each participant will perform sequential biometry scans (BPD, HC, AC, FL) on two distinct pregnant participants scheduled for delivery within 48–72 hours. This sequential design allows for the assessment of "Sensitivity to Change". The primary efficacy endpoint is the

Mean Absolute Percentage Error (MAPE) of the Estimated Fetal Weight (EFW) compared against the definitive Actual Birth Weight (ABW).

3. Participants:

Participants: Participants will be stratified into four cohorts based on their clinical profile and level of training. Because self-reported cumulative numbers of estimated fetal weight (EFW) scans may be difficult to recall accurately, clinical role and training stage will be used as the primary stratification variables. Estimated procedural volume will be collected separately as a descriptive measure of prior hands-on experience and may be used in secondary analyses as an indicator of technical proficiency.

Category	Estimated lifetime EFW scans	Clinical profile	Profesional experience
Novice	< 50 scans	Medical student (final year) or junior doctors without formal OB ultrasound training	No or minimal prior experience in obstetric ultrasound
Intermediate	51-250 scans	Residents undergoing obstetrics/gynecology training (introductory to mid-level – corresponding to intro, post intro, or H1)	Approximately 0.5–4 years of relevant clinical training

Advanced	251 – 1000 scans	Senior residents (H2 and H3) or specialist in obstetrics and gynecology	Several years of supervised and independent ultrasound practice
Expert	>1000 scans	Specialist obstetricians or fetal medicine experts	>5 years or dedicated obestric ultrasound practice

Table 1: Participant stratification*:* Participants are stratified into three cohorts mainly focusing on the procedural volume.

Recruitment of participants: Participants will be recruited through a combination of hospital-based channels and professional networks. Recruitment will take place at Slagelse Hospital and Rigshospitalet, where clinical staff will be invited via internal department meetings and email communication. Medical students currently undertaking OB/GYN clinical rotations will be invited to participate in the Novice cohort. In addition, the study will be advertised through relevant professional forums, medical platforms, and societies to reach a broader range of clinicians across different levels of experience.

The Scan Subjects**:** Pregnant women will be recruited from the induction of labor registry at Rigshospitalet and Slagelse Hospital. To minimize the scan-to-delivery interval and ensure the highest fidelity between Estimated Fetal Weight (EFW) and Actual Birthweight (ABW). The women are planned to give birth within 48-72 hours.

Pregnant women inclusion criteria:

- Pre pregnancy BMI < 40
- Singelton pregnancy
- GA ≥ 37+0 at time of induction
- Intact membranes (to ensure consistent amniotic fluid index)

Exclusion criteria:

- Major fetal anatomical anomaly
- oligohydramnios
- CPR ratio < 2.5th percentile

Recruitment of scan subjects: Recruitment will take place at the induction-of-labor unit at Rigshospitalet and Slagelse hospital. Eligible pregnant women will be approached on the phone 1-2 days prior for induction and informed about the study and have 1-2 days to reply. The day of induction will be the day the scans will take place.

Pre-study clinical Validation and Standardization: To ensure patient safety, every pregnant participant will undergo a standardized baseline assessment by the Principal Investigator (PI) prior to the study intervention. The PI will conduct an initial ultrasound examination to verify fetal well-being. This screening includes amniotic fluid assessment and fetal doppler.

The PI, acting as the Reference Expert, will perform a baseline Estimated Fetal Weight (EFW). This measurement will strictly adhere to the ISUOG Practice Guidelines for the performance of the mid-trimester and third-trimester scan. This reference measurement serves as a secondary point of comparisons alongside the ABW.

To account for variation in technical scanning conditions, the primary analysis will adjust for prespecified patient- and examination-related factors expected to influence ultrasound image acquisition and fetal biometric assessment. These investigator-defined adjustment covariates include maternal pre-pregnancy BMI, fetal position, placental lie, amniotic fluid volume, abdominal scarring, and gestational age. These variables (see table 2.) will be included in the linear mixed-effects model to reduce the risk that increased estimation error is incorrectly attributed to the operator or the AI-assisted method when the examination is technically challenging due to patient-related factors.

Factor	*Low difficulty*	*Moderate difficulty*	*High difficulty*
Maternal pre pregnancy BMI	<25	25-34	>35

Placental lie	Posterior	Lateral	Anterior
Fetal position	Optimal (OA)	Lateral	Deeply engaged/spine up
Amniotic fluid	Normal/polyhydram-nios	Slightly reduced	Oligohydamnios
Abdominal scaring	None	Minor/laparoscopic	Significant shadowing

Table 2: Classification of patient and examination characteristics according to expected level of scanning difficulty, used as adjustment covariates.

4. Intervention:

The AI System: The intervention utilizes a Progressive Concept Bottleneck Model (PCBM) integrated into the ultrasound workflow. Unlike black-box systems, the PCBM provides transparent, landmark and quality-based feedback.

Interface (The Feedback): Participants in the intervention arm will view a real-time overlay of the ultrasound on a screen. The system provides:

- *Landmark Indicators*: Specific visual markers identifying the presence/absence of anatomical landmarks (e.g., *Cavum Septi Pellucidi*, thalamus, umbilical vein, stomach bubble and femur outline).
- *Quality Scoring:* A real-time "traffic light" system (Green = good/yellow=acceptable/Red=poor) indicating when the current view reaches the threshold for a "Standard Plane.

Human-AI Interaction: The intervention is designed as a "Human-in-the-Loop" system.

1 The participant is instructed to scan until the AI signals a "Green" quality score.
2 Once the image is frozen, the AI feedback can also be used.
3 Clinician Verification: Participants evaluates and verify the AI's suggestions.

***AI Model Specification**:* The AI standard plane detection used in this study is based on the Progressive Concept Bottleneck Model (PCBM), a hierarchical variant of the concept bottleneck model [3]. The model is designed to mirror the stepwise decision-making process of experts, by first extracting visual concepts through image segmentation and then applying property concepts that are directly relevant to scan quality assessment.

While conventional 'black-box' AI systems provide binary feedback without revealing the underlying rationale—potentially limiting clinical interpretability and trust—the PCBM offers an explainable AI approach. It provides transparent, clinically meaningful feedback by identifying the presence or absence of specific anatomical landmarks (Figure 1). By highlighting these features, PCBM supports clinical decision-making through active guidance rather than simple classification.

Technical Deployment and Integration: The PCBM was deployed using a bedside-compatible setup to evaluate its clinical usability in a real-world obstetric setting[4] Ultrasound examinations were performed on GE Healthcare E8-10 systems. The ultrasound video output was routed via HDMI to a mini PC (Intel NUC) running the AI software. The generated feedback was transmitted wirelessly to a tablet positioned at the bedside.

***Workflow Standardization*:** To isolate the impact of the AI intervention and minimize dual-screen distraction— the primary ultrasound machine monitor will be blinded during the procedure. Participants will perform the scan relying exclusively on the tablet interface for anatomical visualization.

5. **The study Procedure and workflow:** The experimental session is divided into three phases: Introduction, Intervention and Evaluation:

Phase 1: *Standardized Introduction*:

- *Participant Baseline*: All participants complete a baseline questionnaire detailing their clinical background, years of practice, and lifetime procedural volume.

- *Stratified Randomization*: Participants are randomized (1:1) to either the AI-Intervention Arm or the Standard Care Control Arm. Randomization is stratified by expertise (Novice, Intermediate, or Expert) to ensure balanced cohorts.
- *Standardized Briefing*: To minimize technical bias, participants watch a briefing video on the specific ultrasound equipment and the tablet interface.
- *AI-Intervention Group Only*: This group receives a dedicated tutorial on the AI-based feedback system, including how to interpret automated calipers and manual overrides.
- *Proficiency Check*: Induction concludes with a supervised hands-on practice run and a Q&A session to ensure all participants can navigate the software interface.

Phase 2: *Intervention*: Each participant performs an EFW in two different patients. The workflow follows a two-step process:

1. *Standardized Image Acquisition*: Participants must obtain the three standard anatomical planes required for EFW: Head: Biparietal Diameter (BPD) and Occipitofrontal Diameter (OFD). Abdomen: Transverse Abdominal Diameter (TAD) and Anterior-Posterior Abdominal Diameter (APAD). Femur: Femur Length (FL).
1. *Cursor placement and intervention*: After selecting the optimal images, the participants will measures the biometry:

- *Control group*: Manual cursor placement on the ultrasound machine.
- *AI group*: The PCBM software can be used on the frozen saved images. Participants are required to manually adjust and place the biometric measurement on the ultrasound machine.

2. *Time limit*: A maximum of 30 minutes is permitted per patient for live ultrasound acquisition (probe to skin time). Participants are instructed to complete the required biometry planes as efficiently as possible to reflect real-world clinical workflow. The participants are permitted to exceed the 30-minute limit to perform stationary cursor placement and measurement on the captured freeze images.

Phase 3: Evaluation: Immediately following each scan, participants complete questionnaires

a) Cognitive load (NASA-TLX):

b) Implementation Feedback: System usability will be assessed using the System Usability Scale (SUS), a validated 10-item questionnaire measuring perceived usability and user satisfaction. Scores range from 0 to 100, with higher scores indicating better usability.

6. Outcomes:

Primary outcome:

The primary objective is to evaluate the sensitivity to change in ultrasound measurement accuracy when using AI-feedback compared to standard scanning. We define this as the ability of the AI-system to significantly reduce the measurement error within the same operator class.

- *Primary metric:* Accuracy is measured via the Absolute Percentage Error (APE), calculated as: APE ((EFWscan – BW)/BW) * 100
  *BW:Birth Weight.
- *Primary Comparisons*: The sensitivity to change is analyzed as the difference in the Absolute Percentage Error (APE) between the AI intervention and the Control Group.

Secondary outcome:

*Measurement Deviation from Expert Baseline***:** To assess the precision of the operator relative to a clinical gold standard, the EFW from the study scan will be compared to the baseline EFW obtained by the Principal Investigator (PI) during the initial screening. This "expert-participant delta" measures the AI's ability to reduce inter-observer variability and helps determine if the intervention allows less experienced operators to achieve results closer to expert-level measurements.

*Procedural Efficacy***:** Procedural efficacy is evaluated via total scan duration, measured in seconds from the moment of initial probe-to-skin contact until the final biometry plane is captured. This metric determines whether the AI-intervention may interfere with the length with the scan.

Image Quality: Image quality is assessed using a 16-point score based on the Salomon Criteria. A blinded expert retrospectively evaluates the four standard biometry planes (BPD, HC, AC, and FL) used for the EFW calculation. Each plane is scored based on the visualization of mandatory anatomical landmarks and the accuracy of caliper placement, ensuring that the fetal weight estimation is derived from technically optimized images.

Cephalic	**Abdominal**	**Femoral bone**
1. Symmetrical plane. 2. Plane showing the thalami. 3. Plane showing the cavum septi pellucidi. 4. Cerebellum is not visible. 5. Head plane is occupying more than half of the total image size. 6. Calipers and dotted ellipse placed correctly. - Biparental diameter (½ point) - Occipito-Pariental diameter (½ point)	1. Symmetrical plane. 2. Plane showing the stomach bubble. 3. Plane showing: - the portal sinus (½ point) - Vena umbilicus (½ point) 4. Kidneys are not visible. 5. Abdominal plane. occupying more than half of the total image size. 6. Calipers and dotted ellipse placed correctly. - Antero-Posterior Abdomial Diameter (½ point) - Transverse Abdominal Diameter (½ point)	1. The ends of the bone are clearly visible. 2. < 45 degree angle of the femoral bone and the horizontal plane. 3. Femoral plane is occupying more than half of the total image size. 4. Calipers placed correctly.

Table 3: Modified Salomon criteria. Scoring system used for the objective evaluation of ultrasound image quality. One point was awarded for each fulfilled criterion, with a maximum possible score of 4-6 per standard plane.

Cognitive Load: Cognitive load and user experience are quantified using the NASA-Task Load Index (NASA-TLX). Immediately following the procedure, the operator performs a sub-

jective self-assessment across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. This measures the psychological impact of the AI-interface on the operator's perceived workload.

7. Randomization and allocation:

Randomization: Participants will be randomly assigned 1:1 to either AI-intervention or control arm.

Allocation method: Randomization is performed within each expertise level (novice, intermediate, advanced, and expert) using block randomization with a block size of four. This is to ensure a continuous balance of participants across both study arms even in the event of incomplete recruitment within a cohort.

Concealment: The allocation sequence is computer-generated by R and managed by using RedCAP. The sequence is concealed from participants until they are included in the study.

8. Blinding:

Participants: Unblinded, as the AI feedback is a visible part of the intervention interface.

Expert evaluators: Fully blinded to both the participant's expertise level and their group allocation (AI vs. Control).

9. Power calculation:

The sample size was determined by effect sizes reported in previous studies of ultrasound biometry proficiency. Prior data [5] demonstrated substantial reductions in measurement error following structured intervention, corresponding to a large standardized effect size (Cohen's d ≈ 1.0). As the present study evaluates real-time AI-assisted scanning under routine clinical

conditions rather than a structured training setting, some attenuation of effect is expected. Therefore, a conservative but clinically meaningful effect size in the moderate-to-large range was assumed, corresponding to Cohen's d = 0.80 for the primary outcome.

With a two-sided significance level of 0.05 and 80% power, the confirmatory sample size was set at 72 participants, corresponding to 24 participants in each of three predefined experience strata: Novice, Intermediate, and Advanced. Due to the limited number of fetal medicine experts available nationally, recruitment of 24 participants in the Expert stratum is not feasible. Therefore, the Expert group will be included on an exploratory basis only, and will not contribute to the formal sample size calculation. Within each stratum, participants will be randomized 1:1 to AI-assisted or manual scanning.

The primary outcome will be analyzed using a regression-based model linear mixed-effects model (LMM) comparing AI-assisted and manual fetal biometry, while adjusting for experience stratum and prespecified covariates (e.g., gestational age and patient difficulty).

10. Statistical Analysis:

Fetal weight will be estimated using the Hadlock formula (BPD, OFD, APAD, TAD, and FL).

Primary outcome: The primary objective is to evaluate the sensitivity to change in measurement accuracy, defined as the impact of AI intervention on the Mean Absolute Percentage Error (MAPE) across different expertise levels.

A Linear Mixed-Effects Model (LMM) will be used as the statistical model due to repeated measures. The Absolute Percentage Error (APE) of each individual scan serves as the dependent variable. APE is log-transformed if the residuals exhibits significant skewness.

Fixed effects will include intervention group (AI-assisted vs. control) and operator experience level. The model will adjust for prespecified patient- and examination-related covariates expected to influence scanning conditions, including maternal pre-pregnancy BMI, fetal position, placental lie, amniotic fluid volume, abdominal scarring, and gestational age. A random

intercept for operator will be included to account for clustering of repeated measurements and between-operator variability.

Secondary outcomes:

Secondary endpoints will be analyzed using a consistent LMM or ANCOVA framework to maintain statistical coherence across the study:

- *Procedural Efficacy*: Scan duration (seconds) will be modeled as the dependent variable to assess the temporal impact of the AI-feedback.
- *Image Quality*: The 16-point Salomon Quality Score will be modeled to determine the AI's effect on anatomical plane optimization.
- *Cognitive and Physiological Load*: Both the NASA-TLX (subjective) and GSR (objective) data will be modeled as dependent variables. These analyses will determine if the AI-intervention significantly alters the mental effort or autonomic stress response during the procedure.
- *Measurement Deviation*: The absolute difference between the participant's EFW and the PI's expert baseline EFW will be modeled to assess if AI reduces inter-observer variability.
- *Determination of the Experience Threshold for AI-Mediated Accuracy Gains*: To identify the specific 'cutoff' in clinical experience (measured in years of practice and total lifetime scans) where the transition occurs from significant AI-dependency to manual proficiency. This will be analyzed by calculating the Delta Accuracy (Manual EFW Error minus AI-Assisted EFW Error). A Receiver Operating Characteristic (ROC) curve or a Spline Regression analysis will be used to identify the cutoff point where prior experience significantly diminishes the added value of AI support.

Appendix

1. Sweller, J., *CHAPTER TWO - Cognitive Load Theory*, in *Psychology of Learning and Motivation*, J.P. Mestre and B.H. Ross, Editors. 2011, Academic Press. p. 37-76.
2. Slava Kalyuga, P.A., Paul Chandler, John Sweller, *The Expertise Reversal Effect.* Educational Psychologist, 2011. **38(1)**: p. 23-31.
3. Lin, M.F., A.; Bashir, Z.; Tolsgaard, M. G.; Christensen, A. N., *I saw, I conceived, I concluded: Progressive concepts as bottlenecks.* arXiv preprint arXiv:2211.10630, 2022.
4. Wong, C.K., et al. *Deployment of Deep Learning Model in Real World Clinical Setting: A Case Study in Obstetric Ultrasound*. arXiv preprint arXiv:2211.10630, 2024. 1–10 DOI: 10.48550/arXiv.2404.00032.
5. Andreasen, L.A., et al., *Multicenter randomized trial exploring effects of simulation-based ultrasound training on obstetricians' diagnostic accuracy: value for experienced operators.* Ultrasound Obstet Gynecol, 2020. **55**(4): p. 523-529.