NEJM AI

NEJM AI 2025;2(4)
DOI: 10.1056/AIoa2400747

ORIGINAL ARTICLE

AI to Assist in the Fetal Anomaly Ultrasound Scan: A Randomized Controlled Trial

Thomas G. Day, Ph.D.,[1,2,3] Jacqueline Matthew, Ph.D.,[1] Samuel F. Budd, Ph.D.,[1] Alfonso Farruggia, Ph.D.,[1] Lorenzo Venturini, Ph.D.,[1] Robert Wright, Ph.D.,[1] Babak Jamshidi, Ph.D.,[1] Meekai To, M.D.(res),[3] Huazen Ling, Ph.D.,[4] Jonathon Lai, M.D.,[3,5] Min Yi Tan, M.D.(res),[5] Matthew Brown, M.R.C.O.G.,[6] Gavin Guy, M.D.(res),[6] Davide Casagrandi, M.D.,[7,8] Anastasija Arechvo, Ph.D.,[3] Argyro Syngelaki, Ph.D.,[3,9] David Lloyd, Ph.D.,[1,2] Vita Zidere, Dr.Med.,[2,3] Trisha Vigneswaran, M.D.(res),[2] Owen Miller, F.R.A.C.P.,[1,2] Ranjit Akolekar, M.R.C.O.G.,[6] Surabhi Nanda, M.D.,[10] Kypros Nicolaides, M.D.,[3,9] Bernhard Kainz,Ph.D.,[1,11,12] John M. Simpson, M.D.,[1,2,3] Jo V. Hajnal, Ph.D.,[1] and Reza Razavi, Ph.D.[1,2]

Received: July 25, 2024; Revised: December 9, 2024; Accepted: January 19, 2025; Published: March 27, 2025

Abstract

BACKGROUND Artificial intelligence (AI) has shown potential in improving the performance of screening fetal anomaly ultrasound scans. We aimed to assess the effect of AI on fetal ultrasound scanning in terms of diagnostic performance, biometry, scan duration, and sonographer cognitive load.

METHODS This was a randomized, single-center, open-label trial in a large teaching hospital. Pregnant participants with fetal congenital heart disease (CHD) and participants with healthy fetuses were recruited and scanned with both methods. Sonographers were recruited from regional hospitals and were randomly assigned to scan with either the AI tool or the standard method, blinded to the fetal CHD status. For the AI-assisted scans, the AI models identified and saved 13 standard image planes and measured four biometrics (but did not automate CHD diagnosis). The primary outcome was the diagnostic performance of the scan; secondary outcomes were scan duration and sonographer cognitive load, as well as biometry performance.

RESULTS In total, 78 pregnant participants (26 with fetal CHD) and 58 sonographers were recruited. The sensitivity and specificity of the AI-assisted scan in detecting fetal malformation were 88.9% and 98.0%, respectively, with the standard scan achieving 81.5% and 92.2% (differences in proportions 7.4% for sensitivity (97.5% [CI] confidence interval, −15.9 to 30.7%) and 5.9% for specificity (97.5% CI, −3.8 to 15.5%)). AI-assisted scans were shorter in duration than standard scans (median 11.4 minutes vs. 19.7 minutes, 95% CI for mean difference 7.4 to 11.1). Sonographer cognitive load was lower in the AI-assisted group (median National Aeronautics and Space Administration — Task Load Index [NASA TLX] score 35.2 vs. 46.5, 95% CI for mean difference 4.6 to 15.4). For all biometrics, the AI repeatability and reproducibility were superior to manual measurements. No adverse events were noted during the trial.

CONCLUSIONS AI assistance in the routine fetal anomaly ultrasound scan results in significant time savings, and a reduction in sonographer cognitive load, without a reduction

The author affiliations are listed at the end of the article.

Dr. Day can be contacted at thomas.day@kcl.ac.uk or at the School of Biomedical Engineering and Imaging Sciences, Faculty of Life Sciences and Medicine, King's College London, 9th Floor Beckett House, 1 Lambeth Palace Road, London SE1 7EU, UK.

in diagnostic performance. (The study was funded by an NIHR doctoral fellowship [NIHR301448] and others; ISRCTN number, 65824874.)

Background

Congenital malformations are the most common causes of infant mortality in high-income countries such as the United Kingdom and the United States, and are becoming increasingly important worldwide as other causes of child death become less common.[1] Antenatal diagnosis has been shown to reduce postnatal mortality and morbidity for some lesions, may lead to therapeutic intervention in selected cases, and allows parents to make an informed decision about whether or not they wish the pregnancy to continue.[2,3] The mainstay of antenatal diagnosis is the fetal anomaly screening ultrasound scan. In the United Kingdom, the Fetal Anomaly Screening Programme (FASP) stipulates an offer of this scan between 18^{+0} and 20^{+6} weeks' gestation, with the aim of detecting 11 specific fetal conditions.[4]

Fetal anomaly scans have very high rates of uptake, but universal detection of major fetal malformations has not been achieved. In the United Kingdom, only 50.4% of infants requiring congenital heart disease (CHD) surgery have received an antenatal diagnosis, and there is wide regional variation across the country.[5] Artificial intelligence (AI) has been proposed to improve medical task performance, including the fetal anomaly ultrasound scan.[6] Previous studies have described the development of AI models to automate aspects of the scan, such as plane detection and fetal biometry,[7-11] including a pilot study by our group examining prospective real-time use with normal fetuses.[6]

Despite recent interest in AI across medicine, including obstetrics, high-quality prospective clinical trials remain scarce.[12] While many studies report good model performance using retrospective ultrasound data, no prospective randomized trial has examined the real-world effect of AI on fetal anomaly scans, including abnormal fetuses.

We have created a clinical tool that combines AI models for real-time plane detection, image saving, and biometric measurement with live sonographer feedback. By automating these tasks, the tool streamlines workflow, reducing interruptions and potentially improving fetal malformation detection.

This randomized controlled trial examines the effects of this tool in a population including fetuses with known major structural malformations, involving sonographers from a variety of professional backgrounds. We selected CHD as the focus of the study due to its prevalence, high infant mortality, and frequent missed diagnoses.[13,14] Trial outcome measures were sensitivity and specificity for CHD detection, scan duration, sonographer cognitive load, the quality of saved images, and repeatability and reproducibility of automated measurements. To assess diagnostic performance within a feasible sample size, a novel trial design enriched with CHD-affected fetuses was employed. The variation between participants (both pregnant participants and sonographers) was controlled for by having all pregnant participants undergo both AI-assisted and standard scans, with sonographers randomly assigned to each method.

Methods

STUDY DESIGN AND PARTICIPANTS

The PROMETHEUS trial (Prospective tRial of Machine lEarning To Help fEtal Ultrasound Scanning) was a single-center randomized controlled open-label trial of AI-assisted versus standard unassisted fetal anomaly ultrasound scans. The study was designed so that on a given study day, three pregnant participants (two with a fetus with a normal heart and one with a fetus with CHD) were invited to attend the study site (although on some days, fewer than three pregnant participants actually attended), along with two sonographers, who were randomly assigned to perform scans either with or without AI assistance. Each pregnant participant was scanned twice sequentially, once using each method. The scans were research investigations, not intended to perform a clinical purpose, and performed in addition to standard clinical investigations; thus, standard care was not affected by participation in the trial. The trial was conducted in the Clinical Research Facility of a large urban teaching hospital in the United Kingdom. The study protocol was prospectively registered with the ISRCTN registry as number 65824874. The study was conducted in line with Good Clinical Practice, and all staff members involved in the running of the trial were fully trained in these guidelines.

Pregnant participants were recruited from a tertiary center of fetal cardiology, either following a diagnosis of fetal congenital heart disease ("affected group"), or in whom the fetus had been confirmed to have a normal heart structure after detailed fetal echocardiography ("unaffected group"). The unaffected group had been offered detailed fetal cardiac

screening because of a risk factor for CHD, such as family history, maternal diabetes, or drug exposure. Inclusion criteria were either diagnosis of fetal CHD between 12^{+0} and 27^{+6} weeks' gestation (for the affected group) or confirmation of normal fetal cardiac anatomy between 18^{+0} and 27^{+6} weeks' gestation (for the unaffected group), with at least 1 week between CHD diagnosis and recruitment, if present. Exclusion criteria for pregnant participants were any plan for termination of pregnancy; any known fetal extracardiac structural abnormality at the time of recruitment; any known fetal genetic abnormality; multiple pregnancy; refusal of consent; insufficient English-language skills to provide informed consent; or under 18 years of age. Potential participants were contacted by telephone after approval from the clinical specialist nursing team caring for the patient. The research anomaly scans were performed between 18^{+0} and 27^{+6} weeks' gestation. Each pregnant participant attended a single research session. Participants were enrolled on the day of the scan session, with written consent obtained that day.

Sonography professionals were recruited from units within southeast England via email invitation to the sonographer lead at each department, with a request for them to cascade to their staff. The advertisements were also placed in electronic newsletters of professional groups (e.g., the British Medical Ultrasound Society and the Society of Radiographers). The inclusion criterion was the regular independent performance of fetal anomaly screening ultrasound scans as part of their clinical work. Exclusion criteria were any previous involvement in our research projects or refusal of consent. Written consent was obtained from both pregnant participants and sonographers. They could be from any professional background (e.g., radiography, midwifery, nursing, or medical), and all were termed "sonographers" for the purposes of this study.

RANDOM ASSIGNMENT AND MASKING

Random assignment was performed at a sonographer level on the day of the study session. The pair of sonographers attending that study session were randomly assigned such that one performed the scan with AI assistance, and the other performed a standard manual scan. An online tool was used for the random assignment (randomizer.org). This was used to generate a sequence of two unique numbers, 1 or 2, with 1 corresponding to the AI-assisted scan and 2 corresponding to the manual scan. The first number in the sequence was used to assign the scan type to the sonographer with the surname that was first alphabetically, and the second number was used for the other sonographer. The sonographers were blinded to the clinical status of the pregnant participants (i.e., healthy or CHD), and CHD was the focus of the study.

PROCEDURES

The AI models used in this study performed two tasks: (1) the detecting and labeling of standard image planes from a stream of ultrasound video, and (2) automatic fetal biometry. The models were developed using a prospectively acquired dataset of 7309 complete videos of routine anomaly ultrasound scans performed in a single institution. The training and testing of the AI models are described in the Supplementary Appendix, and described more fully in Venturini et al. and Baumgartner et al.[7,15] The standard planes and biometrics used in the study are shown in Table 1, based on the UK Fetal Anomaly Screening Programme (FASP).[4]

The ultrasound machine used was a GE Voluson Expert 22. Our clinical AI tool consisted of a computer (Boxer-8641AI; Aaeon Technology Inc., Taipei, Taiwan) mounted on the ultrasound machine, receiving the stream of ultrasound video via a high-definition multimedia interface (HDMI) connection. The individual images within the video stream

Table 1. Standard Fetal Ultrasound Image Planes and Associated Biometric Measurements Used in the Study.

Anatomical Area and Standard Plane	Associated Biometric
Head and neck	
Brain transventricular	Head circumference Biparietal diameter
Brain cerebellar	–
Face	
Profile	–
Coronal lips	–
Chest	
Four chamber	–
Left ventricular outflow tract (LVOT)	–
Right ventricular outflow tract (RVOT) or three-vessel view (3VV)	–
Three-vessel tracheal (3VT) view	–
Abdomen	
Abdomen	Abdominal circumference
Transverse kidneys	–
Spine	
Coronal spine	–
Sagittal spine	–
Limbs	
Femur	Femur length

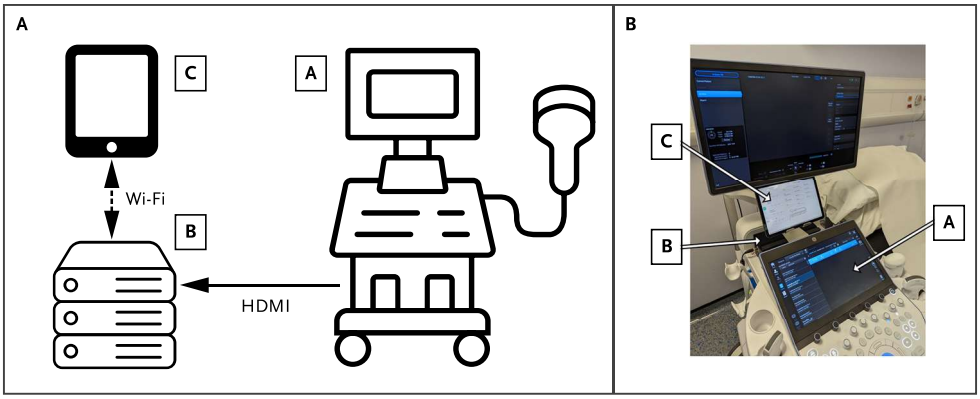

Figure 1. Technical Integration of an Artificial Intelligence Assistance Tool.

Panel A shows a schematic diagram of the study setup. Panel B shows a photograph of the study setup. A represents a standard ultrasound machine; B represents a computer mounted on an ultrasound machine, receiving a video stream of an ultrasound scan via an HDMI cable; and C represents a tablet displaying the outputs of the AI models. The solid black arrow shows a physical cable connection, and the dashed black arrow shows a connection via the Internet. AI denotes artificial intelligence; and HDMI, high-definition multimedia interface.

were analyzed in real time by AI models as described in the Supplementary Appendix, with outputs immediately displayed to the sonographer via a tablet (iPad Air, Fifth Generation; Apple Inc. Cupertino, CA). The tablet was connected to the computer via a Wi-Fi connection. A schematic diagram and photograph of the technical setup are shown in Figure 1.

Sonographers underwent a 15-minute one-to-one training session on the day of the study with an investigator and a written guide and video, as shown in the Supplementary Appendix. They were asked to follow a study-specific scan protocol, as shown in the Supplementary Appendix. Sonographers performing the manual scan were asked to save a single image for each of the 13 standard planes. They were asked to measure each of the four biometric parameters three times and select the best measurement for their report (as per the published guidance).[16] For the sonographers performing the AI-assisted scan, the saving of image planes and measurement of biometrics were performed automatically by the AI tool. Feedback on these processes was shown to the AI-assisted sonographer via the tablet, with example screenshots from the tablet, as shown in Figure 2. During the scan, the current best estimate for each biometric was displayed in real time to the sonographer on a scale indicating the normal values for the given gestation, along with a calculated error bound. Similarly, for each standard plane, a labeled progress bar indicated how many images had been saved. For the AI-assisted scan, after completion, the sonographers were shown a candidate image for each of the standard planes on the tablet. They could either accept this image or choose from a further eight images for each plane to be selected as their final best image (the "image review" stage). The sonographer remained responsible for assessing the imaging data from the scan in order to diagnose fetal malformation. The AI tools did not perform any automated diagnosis of fetal malformation.

After each scan, sonographers completed a written report on a standard laptop using a web-based interface. They were asked to choose a final outcome of either (a) standard follow-up (i.e., usual antenatal care), (b) a repeat scan in a screening department (e.g., owing to limited visibility of an area due to fetal position), or (c) referral to specialist services (e.g., because CHD had been identified). They were asked to make this decision based on the images that they had seen during the scan or image review stage. They also completed survey instruments to measure cognitive load using both the National Aeronautics and Space Administration — Task Load Index (NASA TLX) scale and the Paas scale.[17,18] The NASA-TLX scale was used unweighted, as previously described, and is a multidimensional instrument with six subscales (mental, physical, and temporal demands, and frustration, effort, and performance), which is designed to capture different aspects of cognitive load, each recorded using a visual analog scale and then summed.[19] The Paas scale is a nine-point Likert scale response to the statement, "Please rate your mental effort required to perform the scan."

After the end of the trial, the quality of each saved image was assessed by fetal medicine experts, blinded to the method used to acquire each image, with at least two

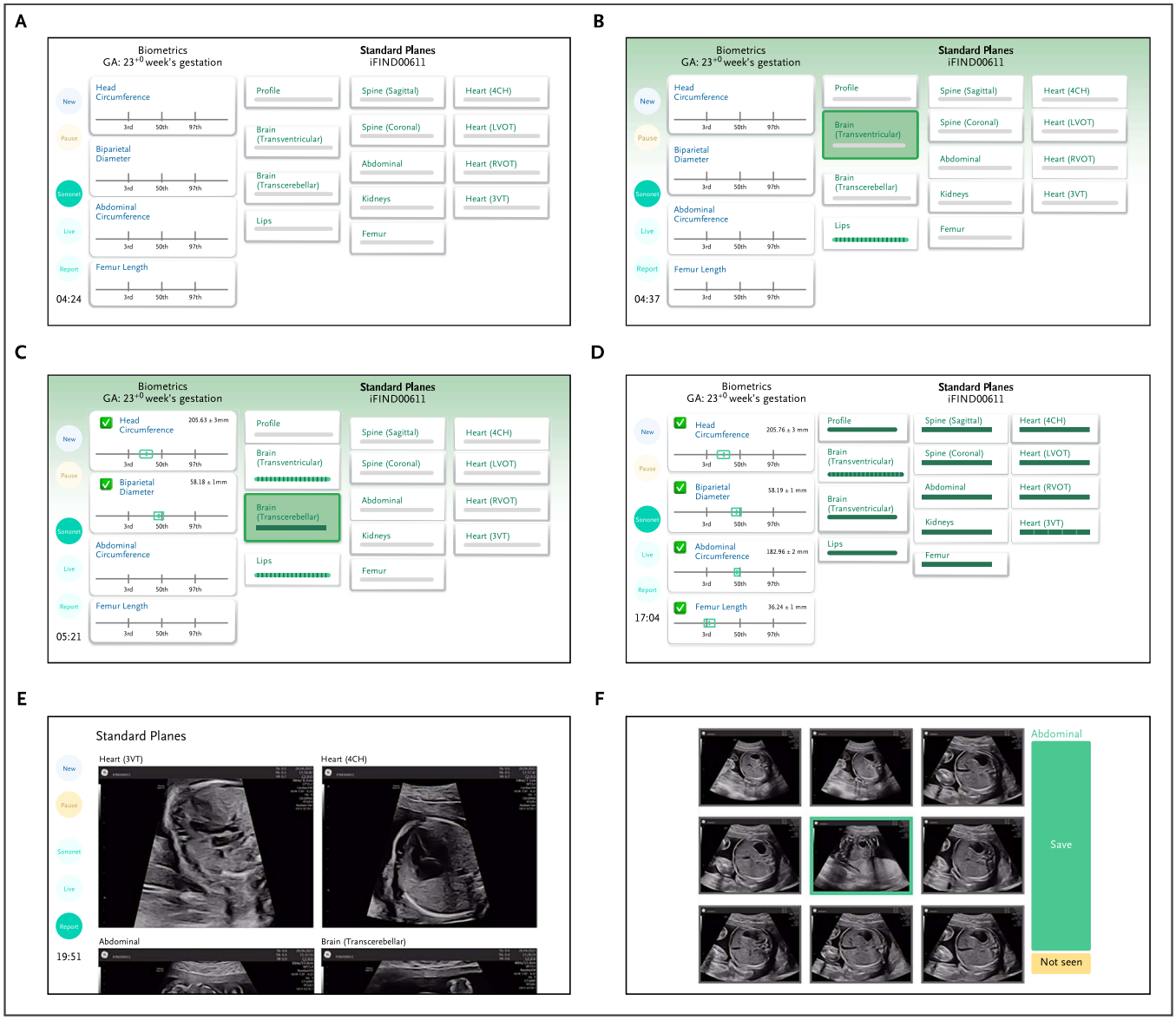

Figure 2. Example Screenshots from the Tablet at Different Time Points of the Examination.

Panel A shows the start of the scan, before any images have been saved. Panel B illustrates that during the scan, the device detected that the brain transventricular view was being imaged, indicated by an overall green color of the upper screen and a green highlight of that view panel. The progress bar below the lips view indicates that some images have also been saved. Panel C illustrates that, during the scan, the brain transcerebellar view is being imaged. The current estimates for the head circumference and biparietal diameter are displayed on the left of the screen on a scale corresponding to the normal range for gestation. Panel D illustrates that at the end of the scan, all biometrics are displayed, and all planes have some saved images. Panel E shows the image review stage, where one candidate image per plane is displayed to the sonographer. Panel F shows that if a plane is selected, a further eight candidate images are displayed for the sonographer to choose from. 3VT denotes three-vessel trachea view; 4CH, four-chamber view; GA, gestational age; LVOT, left ventricular outflow tract; and RVOT, right ventricular outflow tract.

experts scoring each image. Experts agreed to this quality scoring scheme by consensus, with further details in the Supplementary Appendix. This resulted in two metrics for each image: a binary outcome indicating if the image was deemed "clinically acceptable" or not, and a further continuous outcome indicating overall image quality, normalized to a scale from 0 to 1. Further AI models that could automatically discriminate the quality of saved images became available after the trial had commenced. To examine the effectiveness of these, all saved images from the AI

scans were passed through these models, with the top nine images displayed to a research sonographer, as in the live trial. The best quality image from these nine was selected by an experienced research sonographer and assessed using the same quality scoring scheme as the initially selected images. This resulted in a single image being chosen and graded per plane per participant for each of the three methods (manual acquisition, AI acquisition, and AI acquisition with retrospective use of quality models).

OUTCOMES

The primary outcome measures were the sensitivity and specificity of the two methods in detecting CHD. A scan was defined as positive for CHD if at least one of the cardiac views was described as abnormal or not seen in the written report, and the final outcome of the scan was a referral to specialist services. All other scans were defined as negative for CHD. This was compared with the ground truth to classify all scans as true positive, false positive, true negative, or false negative for fetal CHD. If an unaffected fetus was unexpectedly identified as being suspected of having CHD, an urgent repeat specialist fetal echocardiogram was performed on the same day to define whether this was a true- or false-positive finding. Secondary outcome measures were the time taken to complete the scan and report (as measured by a clinical research fellow or research assistant during the study session with a standard stopwatch), the cognitive load of the sonographers (as measured by the administration of survey instruments by a clinical research fellow to sonographers after each scan), and the repeatability and reproducibility of fetal biometrics. All pregnant participants were followed up after delivery to confirm that the antenatal diagnoses were correct. Adverse events of all severities were collected by a clinical research fellow during the scan sessions and at the follow-up contact.

STATISTICAL ANALYSIS

The sample size gave an 80% power to demonstrate non-inferiority of CHD detection sensitivity with a target of 80% and a margin of 25%. The primary outcome measures of sensitivity and specificity of the ultrasound scans were compared by calculating the 97.5% confidence intervals (corrected from 95% intervals as a result of multiple testing, using the Bonferroni method). Confidence intervals for proportions were calculated using the exact Clopper–Pearson method, and for differences in proportions using the Wald method. The secondary outcome measures of scan duration and sonographer cognitive load were compared by calculating 95% confidence intervals for the mean differences between groups. Since these were secondary end points, hypothesis testing was not performed as per journal guidelines. Statistical analysis was performed using SPSS version 29.0.0.0 (IBM Corporation, Armonk, NY). The statistical analysis plan was agreed to before the commencement of data collection.

Manual and AI biometrics were compared using Bland–Altman plots. The three measurements per biometric recorded during the manual scan were used to calculate the repeatability of the manual method (since the mean distance between the maximum and minimum of n observations for a uniform distribution over the interval (a b) is equal to $((n-1)/(n+1)) \times (b-a)$, by using the coefficient two thirds a balance was made between these three-measurement criteria, and the other comparison which had only two measurements). The chosen "best" manual measurement was compared with the final estimate from the AI-assisted scan to compare reproducibility between the two methods. The mean difference between the two methods was also subtracted from the AI measurement, so that the random error could be visualized (i.e., removing any systematic bias). Manual human reproducibility (interobserver variability) was not measured in this trial design, but this has been published previously for three of the four biometrics.[20] Finally, the video recorded during the manual scan was analyzed by the AI model retrospectively to obtain a second AI biometric measurement on the same patient on a sequential scan, so that AI–AI repeatability could be assessed. Figure 3 shows a diagram of how these measurements were compared.

To assess for possible confounders, logistic and quantile regression were performed taking into account differences in the sonographers randomly assigned to each scan method, with primary and secondary outcomes as outputs of these models.

REPORTING

This trial has been reported according to the Consolidated Standards of Reporting Trials (CONSORT)-AI extension (see the Supplementary Appendix).[21]

Results

Figure 4 shows recruitment figures for sonographers, with 58 recruited over the period May 5, 2022, through July 17, 2023. Twenty-nine sonographers were randomly assigned into each scanning method. Baseline characteristics after random assignment are shown in Table 2. Figure 5 shows

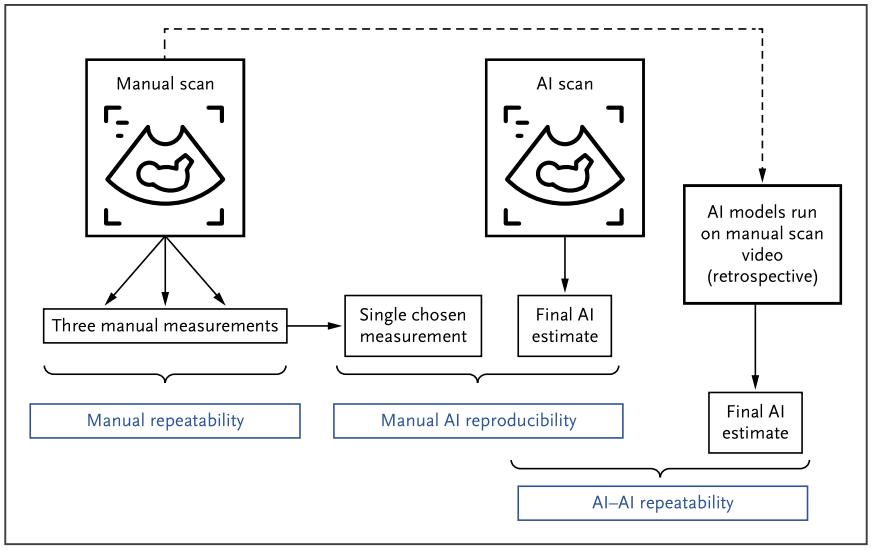

Figure 3. Schematic Diagram Describing Comparison of the Biometric Measurements.

AI denotes artificial intelligence.

recruitment figures for pregnant participants, with 78 recruited over the period November 17, 2022, through August 1, 2023, and included in the final analyses. Baseline characteristics are shown in Table 3, with details of the CHD lesions in cases in Table 4. The study sessions ran from November 15, 2022, to August 8, 2023.

Although 156 paired scans were performed, data were not available from every scan for every outcome measure because of prototype software or hardware failures during the study procedures. This is described in the Supplementary Appendix.

The primary outcome measure for the trial was the diagnostic performance of the scan in detecting fetal CHD. The AI-assisted scan was noninferior to the manual scan both in terms of sensitivity and specificity, with the lower bound of the 97.5% confidence interval (CI) (used rather than 95% CI to correct for multiple testing) crossing zero, but not crossing the prespecified margin of 25%, as shown in Table 5.

Postnatal outcome was available for 73 out of 78 fetuses (93.5%), with five not contactable after birth (all in the unaffected group). All fetuses in the CHD group had a postnatal diagnosis that was concordant with their antenatal cardiac diagnosis. Six fetuses in the group unaffected by CHD had a postnatal echocardiogram due to either a heart murmur on routine postnatal examination or a family history of CHD.

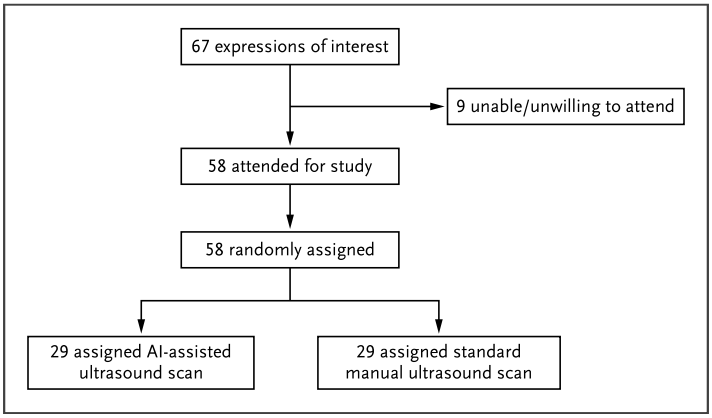

Figure 4. Recruitment Flow Chart for Sonographer Participants.

All sonographers performed the scan method they had been allocated via random assignment. AI denotes artificial intelligence.

Table 2. Baseline Characteristics of Sonographer Participants after Random Assignment.*

Characteristic	AI-Assisted Scan (n=29)	Manual Scan (n=29)
Experience in fetal ultrasound (years)	4 (7)	6 (10)
Professional background		
Radiographer	13 (44.8%)	13 (44.8%)
Nurse/midwife	3 (10.3%)	0
Doctor	13 (44.8%)	16 (55.2%)

*Data are n (%) or median (IQR). AI denotes artificial intelligence; and IQR interquartile range.

Three of these found minor abnormalities that would not be considered detectable on routine antenatal screening (one small secundum atrial septal defect, one very mild pulmonary valve stenosis not requiring treatment, and one subtle hypertrophy of the left ventricle not requiring treatment), so they remained in the unaffected group for the purposes of analysis. The other three were entirely normal.

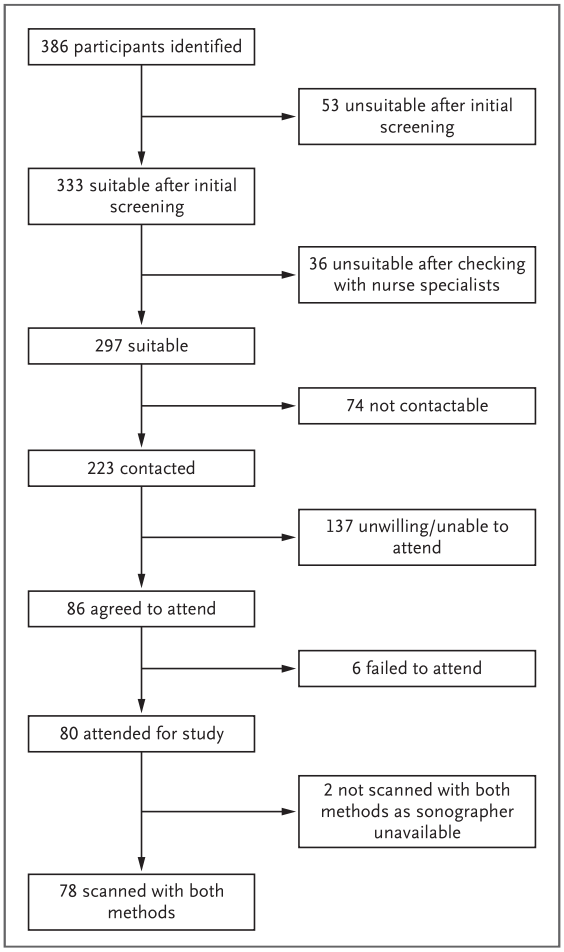

Figure 5. Recruitment Flow Chart for Pregnant Participants.

Table 3. Baseline Characteristics of Pregnant Participants.*

Characteristic	CHD-Affected Group (n=26)	CHD-Unaffected Group (n=54)
Maternal age (years)	32.1 (5.4)	31.2 (6.0)
Gestational age (weeks' gestation)	25.3 (1.9)	24.9 (1.6)
Body mass index	27.9 (5.4)	27.6 (4.6)
Ethnicity†		
White	20 (76.9%)	51 (94.4%)
Black or mixed black	3 (11.5%)	1 (1.9%)
Asian or mixed Asian	1 (3.8%)	2 (3.7%)
Any other	2 (7.7%)	0
Reason for referral to fetal cardiology		
Suspected CHD	24 (92.3%)	3 (5.6%)
Family history of CHD	0	34 (63.0%)
Maternal diabetes	0	5 (9.3%)
Medications	1 (3.8%)	3 (5.6%)
Raised nuchal translucency	1 (3.8%)	6 (11.1%)
Other	0	3 (5.6%)

* Data are n (%) or mean (SD). BMI denotes body mass index (the weight in kilograms divided by the square of the height in meters); CHD, congenital heart disease; and SD, standard deviation.
† Ethnicity was reported by the participants.

Although known fetal extracardiac abnormalities at the time of recruitment was an exclusion criterion, two pregnant participants were included with extracardiac abnormalities as they were identified after recruitment but prior to the study session (one unilateral hydronephrosis, treated conservatively after birth, and one talipes, treated surgically after birth), one of these was also affected by CHD. Both were identified by both scanning methods. In addition, there were four suspected abnormalities identified

Table 4. Details of Fetal CHD Lesions in Cases.*

Fetal CHD Lesion	Number (%)
Right aortic arch	8 (30.8)
Transposition of the great arteries	4 (15.4)
Tetralogy of Fallot	3 (11.5)
Double aortic arch	3 (11.5)
Atrioventricular septal defect	2 (7.7)
Bilateral superior venae cavae	2 (7.7)
Hypoplastic left heart syndrome	1 (3.8)
Double-outlet right ventricle	1 (3.8)
Hypoplastic aortic arch	1 (3.8)
Pulmonary stenosis, right aortic arch, interrupted inferior vena cava	1 (3.8)
Total	26 (100)

*CHD denotes congenital heart disease.

Table 5. Diagnostic Performance of the Two Methods in Detecting Fetal Congenital Heart Disease (CHD-Affected Group, n=26; CHD-Unaffected Group, n=52).*

Diagnostic Metric	AI-Assisted Scan (n=78)	Manual Scan (n=78)	Difference in Proportions (97.5% CI)
True positive (n)	21	20	–
False positive (n)	0	4	–
True negative (n)	52	48	–
False negative (n)	5	6	–
Sensitivity (95% CI)	80.8% (60.6 to 93.4%)	76.9% (56.4 to 91.0%)	3.8% (–18.9 to 26.6%)
Specificity (95% CI)	100% (93.2 to 100%)	92.3% (81.5 to 97.9%)	7.7% (–0.6 to 16.0%)

*AI denotes artificial intelligence; CHD, congenital heart disease; and CI, confidence interval.

during the study scans that were found to be not present on subsequent expert review and after birth (two suspected talipes, one suspected echogenic bowel, and one suspected cleft lip), all were suspected during the AI-assisted scan only. Three of these four had coexisting CHD. Table 6 shows an alternative analysis in which all fetal structural malformations (CHD plus extracardiac anomalies) are considered as the affected group. When analyzed in this way, the AI-assisted scanning method remains noninferior to the manual scanning method in terms of diagnostic performance

The results for scan and reporting duration are shown in Table 7 and Figure 6. The tablet-based image review stage (unique to the AI-assisted scan) was included in the reporting time. The median scan duration was shorter for the AI-assisted scan, with a mean difference in scanning duration of 9.3 minutes (95% CI, 7.4 to 11.1), as shown in Table 7. There was no difference in reporting time between the two groups, with a 95% confidence interval that crossed zero.

The cognitive load of the sonographers was compared between the two groups using both the unweighted NASA-TLX scale and the Paas scale. By both metrics, the sonographers in the AI-assisted scan group reported a lower cognitive load than those in the manual scan group (NASA-TLX median score: 35.3 vs. 46.5 respectively; mean difference 10.0, 95% CI, 4.6 to 15.4); Paas scale score: 5 vs. 6; mean difference 0.95, 95% CI, 0.3 to 1.6). This is shown in Figure 7. For both scan duration analysis and cognitive load, we have not adjusted these secondary end points for multiple testing.

Figure 8 shows the results of the biometric measurements. This analysis was post hoc. The repeatability of the AI

Table 6. Diagnostic Performance of the Two Methods in Detecting All Fetal Structural Malformations (CHD-Affected Group, n=27; CHD-Unaffected Group, n=51).*

Diagnostic Metric	AI-Assisted Scan (n=78)	Manual Scan (n=78)	Difference in Proportions (97.5% CI)
True positive (n)	24	22	–
False positive (n)	1	4	–
True negative (n)	50	47	–
False negative (n)	3	5	–
Sensitivity (95% CI)	88.9% (70.8 to 97.6%)	81.5% (61.9 to 93.7%)	7.4% (–15.9 to 30.7%)
Specificity (95% CI)	98.0% (89.6 to 100%)	92.2% (81.1 to 97.8%)	5.9% (–3.8 to 15.5%)

*AI denotes artificial intelligence; CHD, congenital heart disease; and CI, confidence interval.

Table 7. Scan and Reporting Durations.*

Study Component	AI-Assisted Scan (n=78)	Manual Scan (n=78)	Mean Difference
Scan duration (minutes)	11.4 (3.7)	19.7 (9.6)	9.3 (95% CI, 7.4 to 11.1)
Report duration (minutes)	3.9 (1.7)	4.0 (2.2)	0.1 (95% CI, –0.4 to 0.6)
Combined scan and report duration (minutes)	15.6 (3.9)	24.1 (10.1)	9.4 (95% CI, 7.3 to 11.5)

*Data are medians (IQR). CI denotes confidence interval; and IQR, interquartile range.

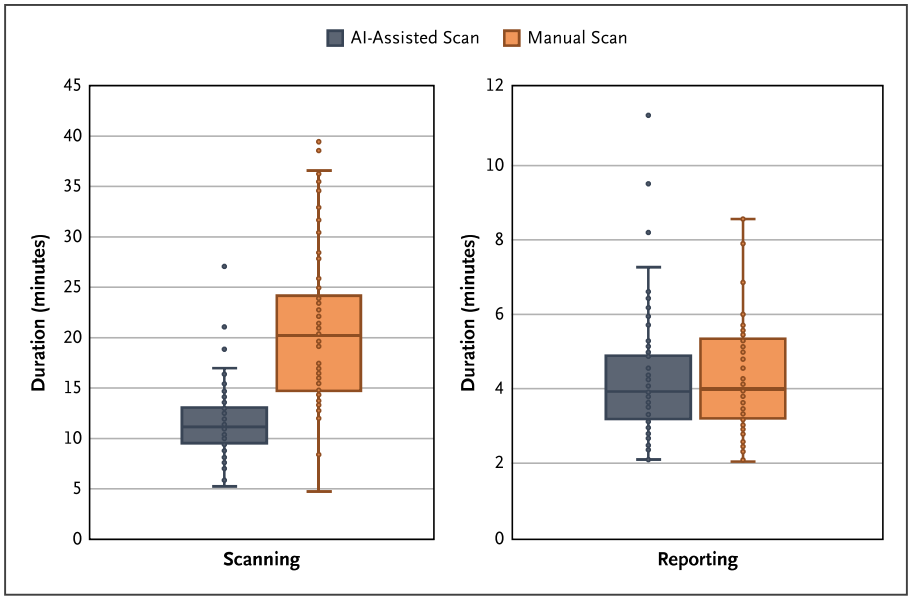

Figure 6. Duration of Scanning and Reporting by Both Methods.

AI denotes artificial intelligence.

measurements (column B) was superior to that of the manual method (column A), as shown by the tighter 95% limits of agreement for repeated measures for all four biometrics. We identified a systematic bias in the reproducibility between AI and manual measurements (i.e., a systematic difference between the two methods, in the absence of a measurable ground truth), from −0.26 (abdominal circumference) to +4.94 mm (head circumference), shown in column C. We did not measure reproducibility between two different human observers, but this has been measured previously[20] and is shown in red in column D (in this column, the systematic bias has been removed by subtracting

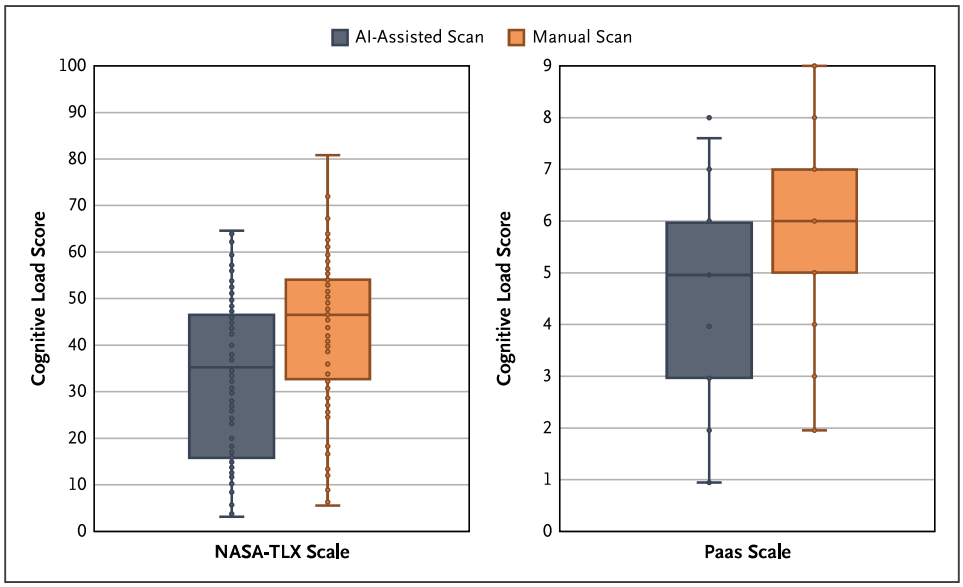

Figure 7. Cognitive Load of the Sonographers Compared between the Two Scanning Methods.

AI denotes artificial intelligence.

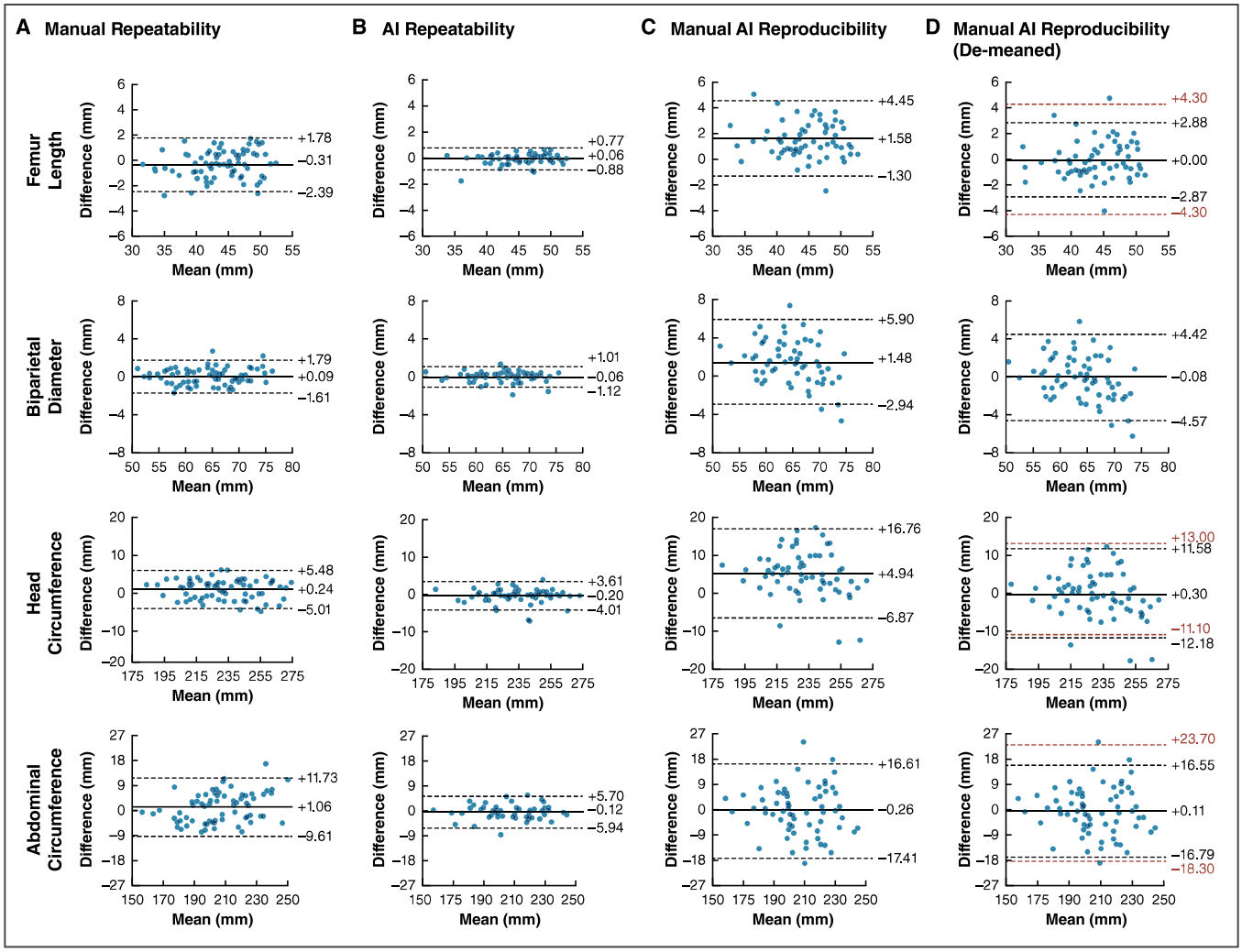

Figure 8. Bland–Altman Plots for the Four Biometric Measurements.

Panel A shows the manual measurement repeatability, based on the three measurements taken during the manual scan. Panel B shows the repeatability of the AI measurement. Panel C shows the reproducibility between the final chosen manual measurement and the measurement from the AI-assisted scan. Panel D, as in Panel C but with the mean AI manual difference subtracted from the AI value, shows only random and not systematic errors, meaning that human interobserver variability (taken from a previous publication by Sarris et al.[20]) can be directly compared (shown in red). The solid lines show the mean difference. The dashed lines show the upper and lower 95% limits of agreement. AI denotes artificial intelligence.

the mean difference to allow direct comparison of the random error between the two groups). The random error seen when comparing AI with manual measurements was less than the random error seen between two humans. However, this estimate of human interobserver variability was made using different data and may not be directly comparable to our findings.

Logistic and quantile regression models were performed to assess for possible confounders. We did not find any significant association between any of the sonographer variables (level of experience and type of professional) and any of the outcome measures, meaning that our overall conclusions were unchanged (data not shown).

Discussion

This randomized controlled trial assesses AI-assisted fetal ultrasound screening, including both normal and abnormal fetuses. We have shown that use of AI can significantly reduce scan duration and sonographer cognitive load while maintaining scan quality in terms of disease detection. Automatically measured fetal biometric measurements

were more repeatable and reproducible compared with human manual measurements. The image quality for some planes was initially inferior using the AI tools, but this was partially ameliorated by the retrospective use of AI models to automatically select the highest-quality images (although further work is needed to improve this for some planes). The trial design enabled a reasonable sample size by enriching for CHD cases, and controlled for variation in both sonographers (through the use of random assignment) and pregnant participants (all were scanned using both methods).

Previous work in the field has focused on the assessment of algorithm performance using retrospective curated test datasets, which may not fully reflect the performance achieved in a real clinical environment.[6-11] A small pilot study by our group has previously suggested a significant time saving with the use of AI, and the present study expands on this by including fetuses with known structural malformations (so that the diagnostic performance of the human–AI team could be assessed), and involving a large cohort of randomly assigned and blinded sonographers.[6] These results are encouraging and suggest that a real clinical benefit may be offered if AI is integrated into current fetal ultrasound screening programs.

On average, sonographers with AI assistance saved around 42% of the scan duration, potentially allowing more time to improve the overall scan experience, such as communicating with the patient or spending more time imaging a particular anatomical area of concern. Shorter scans may also yield health economic benefits by lowering costs. Although this study did not assess where time savings occurred, our previous pilot study analysis found that the time savings corresponded to the periods of time where, in the manual scan, the screen was frozen (i.e., the time taken to measure biometry and save still images).[6] However, it is worth noting that we have not proved in the present study that these time savings would necessarily translate into actual clinical benefit. Larger, real-world trials are required to demonstrate this benefit in clinical practice.

Our findings suggest that the AI-assisted scans were not inferior to the manual scans in terms of diagnostic performance. This is reassuring, as it implies that workflow efficiencies can be achieved without an adverse effect on sensitivity and specificity. As previously demonstrated, specificity, in particular, is extremely important when considering the introduction of AI to screening programs to avoid overwhelming downstream specialist services with false-positive referrals.[22] The fact that specificity remains robustly high in the AI-assisted group offers reassurance that AI tools may prevent this issue, potentially reducing false-positive referrals for CHD. However, this study was only powered to detect large differences in diagnostic performance, with a margin of 25%. There could be smaller clinically relevant differences between the two scanning methods that we were underpowered to detect, so our results should be interpreted with caution. For this reason, we are planning a larger multicenter trial of this technology, which will have a larger study power but will require a sample size of several thousand participants.

AI-measured biometrics were found to be more repeatable and had less random error than manual measurements by different sonographers. We did identify a difference between manual and AI measurements, but accuracy remains unclear without a gold standard. It would be relatively trivial to convert the AI measurement to an equivalent of the manual measurement by subtracting the detected difference, if that were desirable. The AI measurements were based on tens or hundreds of measurements per scan using a Bayesian approach,[15] rather than the traditional approach of measuring just three times (or in many cases, just once). This resulted in a final estimate that had far higher repeatability compared with manual measurements, as well as having the advantage of real-time feedback to the sonographer of the error range around the currently estimated measurement. This method could be applied to many ultrasound-based measurements, even beyond obstetrics. By reducing random human error, we can obtain measurements that are precise, even if the scan is conducted by a different operator. Such measurements are often extremely clinically important, and by reducing variability, we can be more confident about thresholds used to instigate or monitor treatments.

Despite a short training period on a novel system, AI-assisted sonographers still recorded significantly lower cognitive load scores by two different metrics compared with the manual group. Cognitive load is a concept describing how mentally challenging a given task is, and reducing it by taking over specific mundane and/or distracting tasks is a potential mechanism by which the human–AI team performance might exceed that of humans alone.[23] Limited research exists on cognitive load changes after medical interventions and the effect on clinical practice. However, looking at the published distribution of scores across multiple fields, such as air traffic control, the reduction seen here in NASA-TLX score of around 11 points is similar to a reduction from the 50th to the 25th centile, which is likely to correspond to a substantial perceived difference in load.[24] The combination of reduced cognitive load and reduced scan duration shows exciting potential, and one

we hope will be translated into improved fetal ultrasound screening outcomes.

Image quality results were mixed. Initially, AI-acquired images were rated lower than manually acquired ones. However, when an improved AI model was utilized retrospectively to select the highest-quality images from the same recorded examinations, this problem was solved for many of the planes. This indicates that the problem for these planes was not that high-quality images were not saved, it was that the candidate images presented to the sonographers were initially not the subjectively "best" ones out of all the saved images. Our current image quality models have performed well, but some image planes still require further improvement — probably via enlargement of the training set with further labeled images — to match the quality of manually saved images. These quality models could be integrated into the overall clinical tool and used in real time for future studies.

The main limitation of this trial was that we conducted research ultrasound scans, performed in addition to the standard clinical pathway. The sonographers were self-selected, and may not reflect the broader workforce of sonographers in terms of professional background or skill level. We have included multiple professional groups operating as sonographers in the U.K., which may not be representative of all international health systems (e.g., some countries would exclusively use medical doctors to undertake this task). Although they were blinded to the CHD status of each fetus, and even though CHD was the focus of the study, they were aware that a potential malformation was present. Given that each sonographer only scanned a maximum of three participants, they may have been more cautious or thorough compared with their usual clinical practice. Because of the current limitations of fetal ultrasound screening, we could not recruit pregnant participants from the standard screening population as we would not have access to a reliable ground truth. For this reason, we recruited participants who had undergone detailed fetal echocardiography, a procedure that has a much higher sensitivity and specificity than standard ultrasound screening.[25] This means that we only included cases of CHD that had already been detected. How well an AI tool assistance works in an unselected population has not yet been assessed. We also used a single model of an ultrasound machine, meaning that potential domain-shift problems that may be encountered in clinical use have not yet been fully explored. There are also multiple types of CHD included in this study, which will vary in their relative likelihood of antenatal detection.

We have not addressed some broader concerns regarding medical AI in this trial, such as the potential for workforce deskilling by the automation of specific tasks. If sonographers perform fewer manual scans, skills may decline, impacting patient care if AI fails. This is an important issue and will need to be addressed if AI is to be translated to the clinical environment, perhaps by ensuring sonographers still undertake some manual scans intermittently. However, some other broader concerns, such as inattention to anomalies secondary to "automation bias," have been addressed in this study, and our findings are reassuring on this front. Many risks of AI are at least partially mitigated by ensuring that the AI tool and human operator work in partnership, and that the human always retains complete control over the final interpretation of the scan findings. Translation of this study to a real-world screening-level population will be key to fully exploring the utility of AI tools. Given the prevalence of congenital anomalies in the general population, this will likely require a large multisite trial, recruiting participants as they undergo their routine clinical screening ultrasound. Enlargement of the study population in this way would also address the sample size of the present study, which is a limitation in terms of detecting more subtle effects on detection rates. However, the present study was large enough to demonstrate a significant positive effect of AI assistance. We are also working on broadening our range of views that can be automatically captured to include all the views recommended by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG).[26]

Previous work has also explored the use of AI to directly detect anomalies such as CHD on ultrasound images.[22,27-29] The addition of such models to our current AI tool may further improve overall scan performance, but this needs to be carefully assessed due to potential risk. However, such model ensembles may be a powerful way of improving the detection of fetuses with congenital malformations and will be the focus of our future work.

In summary, we have demonstrated that AI assistance for fetal ultrasound screening is safe and effectively reduces scan duration and cognitive load without reducing diagnostic performance. This is one of the relatively few randomly assigned prospective controlled trials of AI in medicine and raises the exciting prospect of future human–AI collaboration in this field.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

The study was funded by a National Institute for Health and Care Research doctoral fellowship (NIHR301448). It was supported by grants from the Wellcome Trust (IEH Award, 102431), by core funding from the Wellcome Trust/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z), and the London AI Centre for Value Based Healthcare via funding from the Office for Life Sciences. T.D. is currently supported by a Clinical Research Excellence Fellowship from King's Health Partners Centre for Translational Medicine. B.K. received funding from the European Research Council (project MIA-NORMAL 101083647).

Patient-level imaging data from the trial and the imaging data used to train the AI models are not available for sharing due to ethical restrictions. The study protocol, patient information sheet, and example consent forms are available on request. The code used to train the AI models is available on request, but the model weights used in the trial are not available.

We thank all the pregnant and sonographer participants of this trial who gave up their time to try and improve fetal ultrasound screening. We also thank the nursing and medical staff of the Evelina Children's Hospital Fetal Cardiology Unit, for their help in recruiting the participants for this study. Finally, we thank Caitlin Giles and Abigail Adeosun, along with the staff of the St Thomas' Hospital Clinical Research Facility, for the smooth running of the trial.

Figures 1 and 3 were created using images from Flaticon.com.

Ethics approval was granted by the London Dulwich Research Ethics Committee (as reference 22/LO/0163.)

Author Affiliations

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London

[2] Department of Congenital Heart Disease, Evelina London Children's Healthcare, Guy's and St Thomas' NHS Foundation Trust, London

[3] Harris Birthright Centre, King's College Hospital, London

[4] Department of Fetal Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London

[5] Department of Fetal Medicine, St Mary's Hospital, Imperial College Healthcare NHS Trust, London

[6] Department of Fetal Medicine, Medway Maritime Hospital, Medway NHS Foundation Trust, Gillingham, UK

[7] Department of Fetal Medicine, University College London Hospitals NHS Foundation Trust, London

[8] Elizabeth Garrett Anderson Institute for Women's Health, University College London, London

[9] Fetal Medicine Foundation, London

[10] Department of Fetal Medicine, Guy's and St Thomas' NHS Foundation Trust, London

[11] Department of Computing, Imperial College London, London

[12] Image Data Exploration and Analysis Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

References

1. Office for National Statistics. Child and infant mortality in England and Wales: 2021. Newport: Office for National Statistics, 2023 (www.ons.gov.uk).

2. Holland BJ, Myers JA, Woods CR. Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise prior to planned neonatal cardiac surgery: a meta-analysis. Ultrasound Obstet Gynecol 2015;45:631-638. DOI: 10.1002/uog.14882.

3. Cortes MS, Chmait RH, Lapa DA, et al. Experience of 300 cases of prenatal fetoscopic open spina bifida repair: report of the International Fetoscopic Neural Tube Defect Repair Consortium. Am J Obstet Gynecol 2021;225:e1-678.e11. DOI: 10.1016/j.ajog.2021.05.044.

4. NHS Screening Programmes. NHS fetal anomaly screening programme handbook. London: HM Government, 2018.

5. National Cardiac Audit Programme. National congenital heart disease audit (NCHDA) 2021 summary report. London: Healthcare Quality Improvement Partnership, 2021.

6. Matthew J, Skelton E, Day TG, et al. Exploring a new paradigm for the fetal anomaly ultrasound scan: artificial intelligence in real time. Prenat Diagn 2022;42:49-59. DOI: 10.1002/pd.6059.

7. Baumgartner CF, Kamnitsas K, Matthew J, et al. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans Med Imaging 2017;36:2204-2215. DOI: 10.1109/TMI.2017.2712367.

8. Cai Y, Sharma H, Chatelain P, Noble JA. Multi-task SonoEyeNet: detection of fetal standardized planes assisted by generated sonographer attention maps. Lect Notes Comput Sci Subser Lect Notes Artif Intell Lect Notes Bioinforma 2018;2018:871-879. DOI: 10.1007/978-3-030-00928-1_98

9. Burgos-Artizzu XP, Coronado-Gutiérrez D, Valenzuela-Alcaraz B, et al. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci Rep 2020;10:1-12. DOI: 10.1038/s41598-019-56847-4.

10. Sinclair M, Baumgartner CFCF, Matthew J, et al. Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS, 2018:714-717.

11. Gao J, Lao Q, Liu P, et al. Anatomically guided cross-domain repair and screening for ultrasound fetal biometry. IEEE J Biomed Health Inform 2023;27:4914-8725. DOI: 10.1109/JBHI.2023.3298096.

12. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomized controlled trials evaluating AI in clinical practice: a scoping evaluation. 2023 September 13 (https://www.medrxiv.org/content/10.1101/2023.09.12.23295381). Preprint.

13. Gilboa SM, Salemi JL, Nembhard WN, Fixler DE, Correa A. Mortality resulting from congenital heart disease among children and adults in the United States, 1999 to 2006. Circulation 2010;122:2254-2263. DOI: 10.1161/CIRCULATIONAHA.110.947002.

14. Aldridge N, Pandya P, Rankin J, et al. Detection rates of a national fetal anomaly screening programme: a national cohort study. Br J Obstet Gynaecol 2023;130:51-58. DOI: 10.1111/1471-0528.17287.

15. Venturini L, Budd S, Farruggia A, et al. Whole-examination AI estimation of fetal biometrics from 20-week ultrasound scans. 2024 January 2 (https://arxiv.org/abs/2401.01201). Preprint.

16. British Medical Ultrasound Society. Professional guidance for fetal growth scans performed after 23 weeks of gestation. London: British Medical Ultrasound Society, 2022.

17. National Aeronautics and Space Administration. NASA task load index. 2020 (https://humansystems.arc.nasa.gov/groups/TLX/).

18. Paas FGWC. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. J Educ Psychol 1992;84:429-434. DOI: 10.1037/0022-0663.84.4.429.

19. Hart SG. NASA-TLX: 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006:50:904-908. DOI: 10.1177/154193120605000909.

20. Sarris I, Ioannou C, Chamberlain P, et al. Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet Gynecol 2012;39:266-273. DOI: 10.1002/uog.10082.

21. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2:e537-e548. DOI: 10.1016/S2589-7500(20)30218-1.

22. Day TG, Budd S, Tan J, et al. Prenatal diagnosis of hypoplastic left heart syndrome on ultrasound using artificial intelligence: how does performance compare to a current screening programme? Prenat Diagn 2023;44:717-724. DOI: 10.1002/pd.6445.

23. Ehrmann DE, Gallant SN, Nagaraj S, et al. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nat Med 2022;28:1331-1333. DOI: 10.1038/s41591-022-01833-z.

24. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. Proc Hum Factors Ergon Soc Annu Meet 2015;59:1727-1731. DOI: 10.1177/1541931215591373.

25. Donofrio MT, Moon-Grady AJ, Hornberger LK, et al. Diagnosis and treatment of fetal cardiac disease: a scientific statement from the American Heart Association. Circulation 2014;129:2183-2242. DOI: 10.1161/01.cir.0000437597.44550.5d.

26. Salomon LJ, Alfirevic Z, Berghella V, et al. ISUOG practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet Gynecol 2022;59:840-856. DOI: 10.1002/uog.24888.

27. Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med 2021;27:882-891. DOI: 10.1038/s41591-021-01342-5.

28. Budd S, Sinclair M, Day T, et al. Detecting hypo-plastic left heart syndrome in fetal ultrasound via disease-specific atlas maps. In: de Bruijne M, Cattin PC, Cotin S, et al., eds., Medical image computing and computer assisted intervention — MICCAI 2021. Cham: Springer International Publishing, 2021:207-217.

29. Tan J, Au A, Meng Q, et al. Automated detection of congenital heart disease in fetal ultrasound screening. In: Hu Y, Licandro R, Noble JA, et al., eds., Medical ultrasound, and preterm, perinatal and paediatric image analysis. ASMUS PIPPI 2020. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020:243-252.