Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Teeranan Pokaprakarn, Juan C. Prieto, Joan T. Price, et al. AI Estimation of Gestational Age from Blind Ultrasound Sweeps in Low-Resource Settings. NEJM Evidence. DOI: 10.1056/EVIDoa2100058.

Supplementary Appendix

Table of Contents

Section 1: Technical Methods for the Deep Learning Model

Model Architecture

The model architecture is graphically illustrated below (Figure S1) and consists of two modules: Feature Extraction Module and Weighted Average Attention Module.

Feature Extraction Module

During training, each frame is processed using ResNet-50[1] architecture initialized with weights trained on the ImageNet[2] data set, this step yields a feature vector of size 2048 for each frame. The extracted features are then analyzed via our Weighted Average Attention (WAA) Module described in the following section. While a pre-trained network is used for feature extraction,[3-6] the weights in ResNet-50 are fine-tuned together with the other parameters in the model during training.

Weighted Average Attention Module

The functional form of our attention module is motivated by the additive Bahdanau attention.[7] Our Weighted Average Attention (WAA) Module has 3 trainable parameters V, W, and Q as defined here:

$$\boldsymbol{w_t} = f(\boldsymbol{x_t}) = \sigma\left(V\left(tanh\left(W(\boldsymbol{x_t})\right)\right)\right) \quad (1)$$

$$\boldsymbol{s_t} = \frac{\boldsymbol{w_t}}{\sum_{u=1}^{N} \boldsymbol{w_u}} \quad (2)$$

$$\boldsymbol{a} = \sum_{t=1}^{N} \boldsymbol{s_t} \cdot Q(\boldsymbol{x_t}) \quad (3)$$

where $\boldsymbol{x_t}$ is the output features from the feature extraction module (t is the frame index, N is the total number of frames). The attention module comprises W which is a dense layer followed by the hyperbolic tangent activation function and then another dense layer V to map to a single scalar value $\boldsymbol{w_t}$ between zero and one for each frame with a sigmoid function. Equation (2) computes a weighted score $\boldsymbol{s_t}$ on the time dimension so that $\sum_t^N \boldsymbol{s_t} = 1$.

The parameters of the weighted average attention module are jointly trained with the other parts of the model. The attention mechanism described above allows the model to focus on frames of the input sequence that contain fetal structures and maximize the gestational age prediction power. The output of this attention module is a weighted sum of the features from the input frames and computed as shown in Equation (3) where Q is another dense layer which reduces the dimension of the feature vector $\boldsymbol{x_t}$ from 2048 to 128. The weighted sum computation allows arbitrary sequence length and enables our model to make predictions based on a single or multiple frames.

Finally, a single linear layer takes $\boldsymbol{a}$ as input and output the gestational age estimate.

Training Procedure

In a pre-processing step, the blind sweeps are re-sampled to a common space with spacing of 0.75 mm/pixel and image dimensions of 256x256 pixels (i.e., physical size of 192x192 mm). To ensure that our network is learning solely from ultrasound image features, we mask and crop text that could bias learning. The number of individual frames comprising each blind sweep cineloop is left unchanged.

Sonographers were instructed to collect blind sweep cineloop videos that were approximately 10 seconds in length. In practice, this length varied as did the corresponding number of frames (median = 180 frames; maximum > 600 frames). Because training with all available frames in these longer sequences is computationally intensive, we select 50 frames at random from all those available in a given sweep. We also apply the following additional processing to each frame from the blind sweep: padding to 288x288 image, and then 256x256 random cropping. The images are then loaded into the range [0, 1] and standard normalization (mean subtraction and standard deviation division by channel; both from ImageNet) is applied as required for the input of ResNet-50.

We use the adaptive moment estimation (ADAM) optimization algorithm[8] with a learning rate of 10^{-4} and a batch size of 24 blind sweeps. Each training epoch contains 5000 batches in which weighted sampling is used to sample blind sweeps into each batch to handle the imbalanced distribution of gestational age. The loss function is Mean Absolute Error (MAE). We use an early stopping procedure[9] and track the MAE of the studies in the tuning set, where for a given study all frames are concatenated to form a single sequence and used for prediction. When the tuning set MAE does not improve over 10 epochs, we stop the training and save the best model as evaluated by the tuning set MAE.

Inference

At inference, all available blind sweeps for a study are concatenated and all available frames are used for prediction. Note that this differs from the training procedures, where only a random subset of 50 frames per blind sweep were used. No data augmentation is applied and the full concatenated sequence of 256x256 cropped frames is provided to the model. In a manner similar to that described above for training, each frame is loaded into the range [0, 1] and normalization is applied.

Section 2: Additional Statistical Methods

Calculation of Standard Errors (SE) for Root Mean Square Error (RMSE)

Let $U_i, i = 1, \dots, n$ be n independent errors.
Denote RMSE as $\hat{\sigma}_n$, where $\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n U_i^2$.
Define

$$\hat{\tau}_n^2 = \frac{\sum_{i=1}^n \left[U_i^4 - \hat{\sigma}_n^4\right]}{(4n\hat{\sigma}_n^2)}$$

Then the SE for the RMSE is $\hat{\tau}_n/\sqrt{n}$ based on the delta method.

Calculation of Standard Errors (SE) and Confidence Interval for Difference in Root Mean Square Error (RMSE)

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be the errors for the model (X) and for the expert (Y). Let $\hat{\sigma}_n^2 = \mathrm{P}_n X^2$ and $\hat{\tau}_n^2 = \mathrm{P}_n Y^2$, where P_n is the empirical process (i.e., $\mathrm{P}_n f(X) = n^{-1} \sum_{i=1}^n f(X_i)$). Also, denote $\sigma_0^2 = PX^2$ and $\tau_0^2 = PY^2$, where P is the expectation.

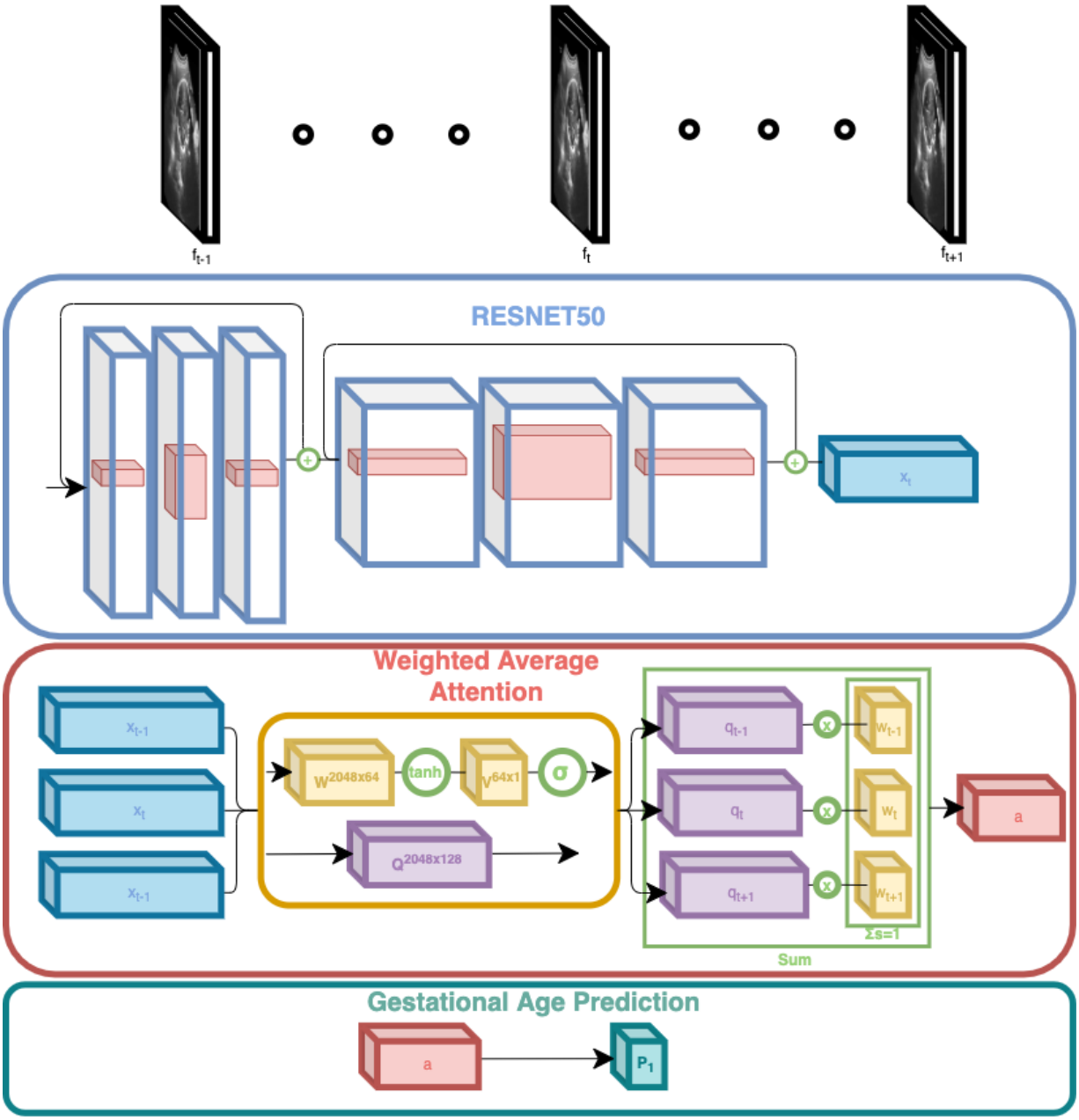
We can use the Taylor expansion, and the fact that the derivative of $\sqrt{u}$ is $u^{-1/2}/2$, to obtain that $\sqrt{n}\left(\sqrt{\hat{\sigma}_n^2} - \sigma_0\right) = \sqrt{n}(\hat{\sigma}_n^2 - \sigma_0^2)/(2\sigma_0) + o_P(1)$. Similarly, we can verify that $\sqrt{n}\left(\sqrt{\hat{\tau}_n^2} - \tau_0\right) = \sqrt{n}(\hat{\tau}_n^2 - \tau_0^2)/(2\tau_0) + o_P(1)$. Now, letting $IF(X,Y) = (X^2 - \sigma_0^2)/(2\sigma_0) - (Y^2 - \tau_0^2)/(2\tau_0)$, we have that $D_n = \sqrt{\hat{\sigma}_n^2} - \sqrt{\hat{\tau}_n^2} - \sigma_0 + \tau_0 = n^{-1/2}\mathrm{P}_n IF(X,Y) + o_P\left(n^{-1/2}\right)$.
This means that the true variance of D_n equals $n^{-1}P\left(IF(X,Y)\right)^2 + o(1)$.

Let $\widehat{IF}(X,Y) = (X^2 - \hat{\sigma}_n^2)/(2\hat{\sigma}_n) - (Y^2 - \hat{\tau}_n^2)/(2\hat{\tau}_n)$.
$n^{-1} \sum_{i=1}^n \left[\widehat{IF}(X_i, Y_i)\right]^2$ is consistent for $P[IF(X,Y)]^2$, and thus we can consistently estimate the the SE of the difference between the RMSEs with

$$n^{-1/2} \sqrt{n^{-1} \sum_{i=1}^n \left[\widehat{IF}(X_i, Y_i)\right]^2}$$

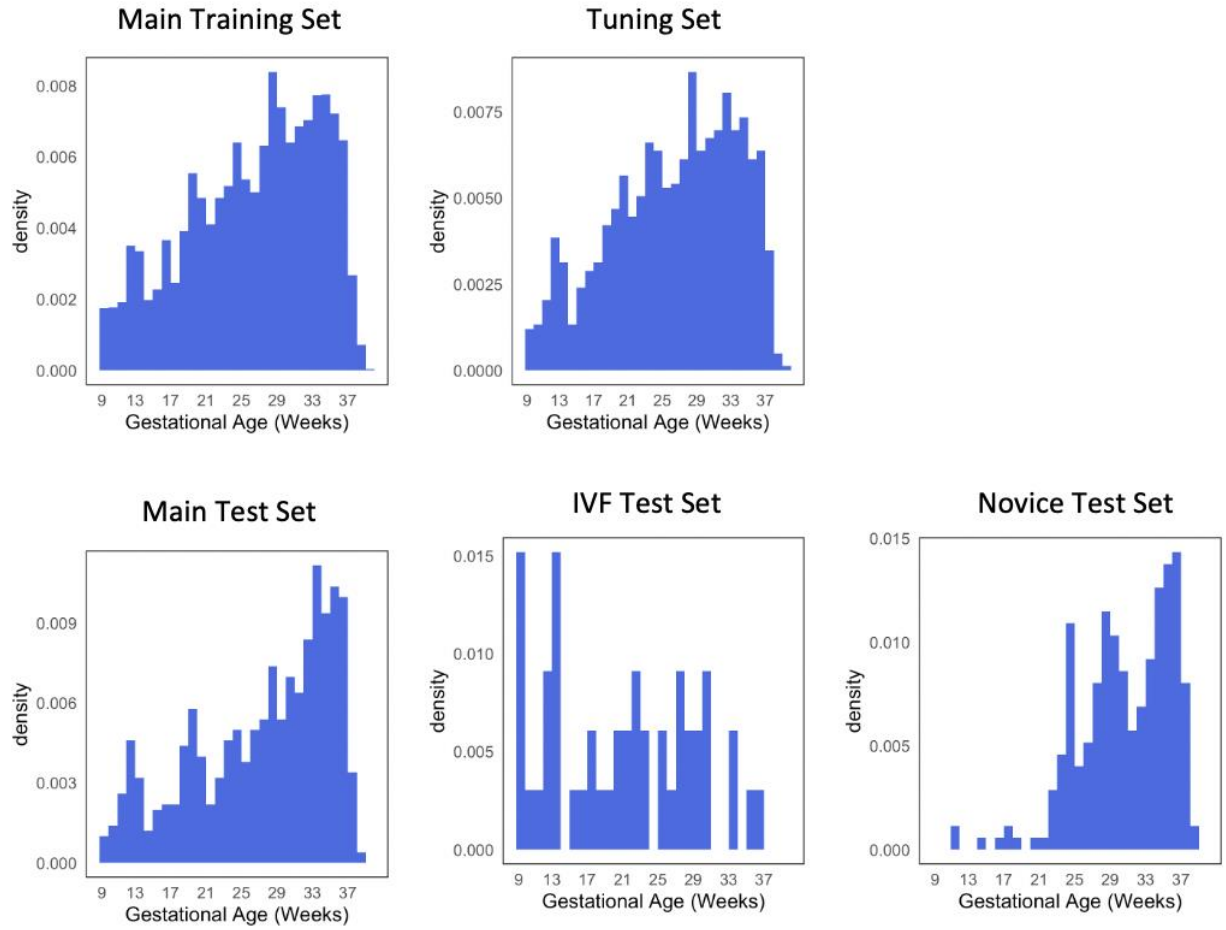
Section 3: Supplementary Figures

Figure S1: DL model architecture



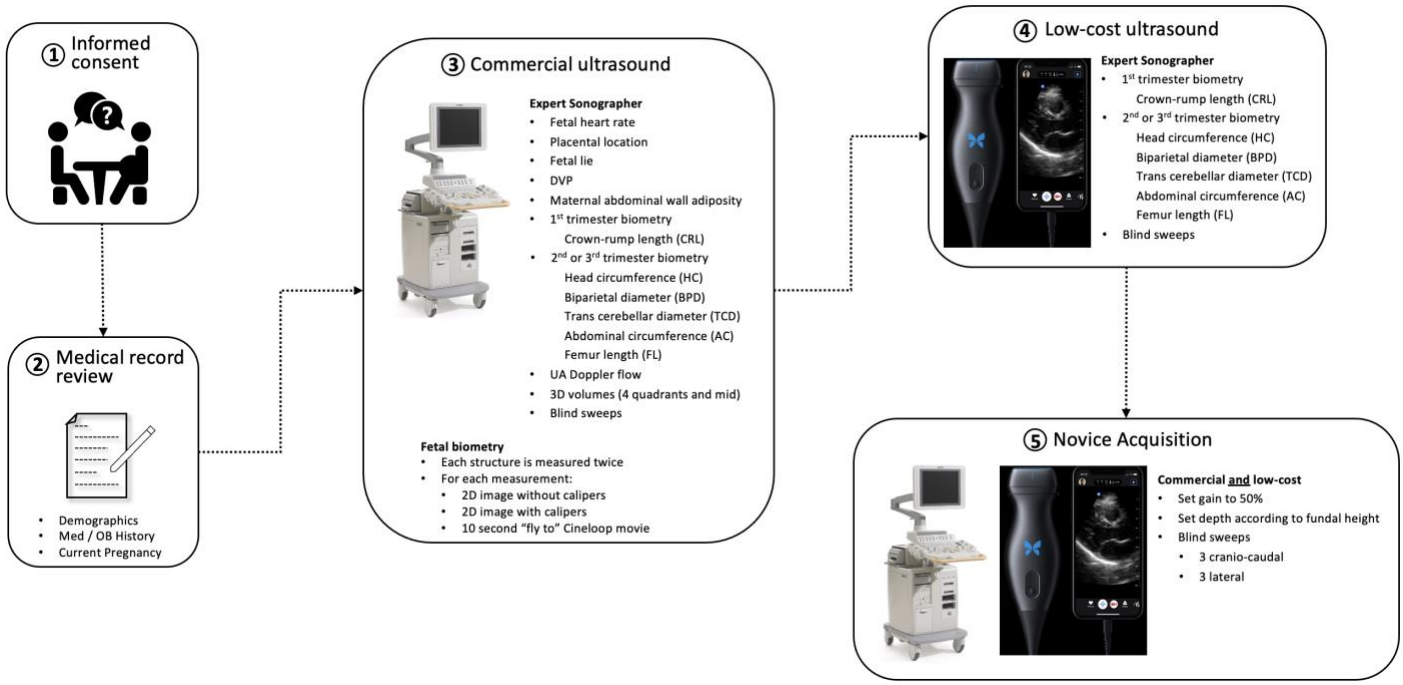
Feature extraction from each frame using a RESNET50 architecture. During training, the RESNET50 is initialized using weights pre-trained on the ImageNet dataset. Each feature vector xt with dimension 2048 is used as input to our Weighted Average Attention Layer where a score w [0-1] is assigned to a frame using two fully connected layers (W, V). The dimension of each feature vector is reduced to 128 using a fully connected layer (Q). We use the scores to compute a weighted sum vector a which summarizes any input sequence with variable number of frames. Finally, a fully connected layer (P) predicts the gestational age.

Figure S2: "Ground Truth" Gestational Age Distribution of the Training and Testing Datasets



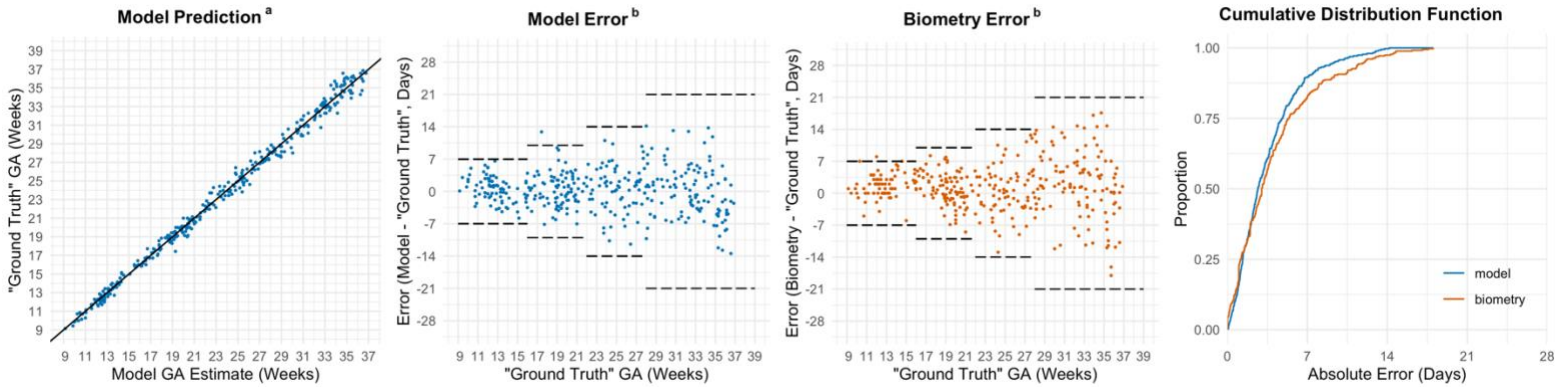
In Zambia, "ground truth" gestational age is defined by the first ultrasound. In North Carolina, it is defined by an algorithm incorporating both the last menstrual period and the first ultrasound,[10] unless the pregnancy was conceived by *in vitro* fertilization (in which case the exact date of fertilization is known).

Figure S3: Overview of FAMLI Protocol Clinical Data Collection



Graphical representation of a participant visit and ultrasound data collection in the FAMLI Study. Step 5 (novice acquisition) began in June 2020 at the Zambia sites only.

Figure S4: Gestational Age Estimation of Deep Learning Model Compared to Trained Sonographer in the Subset of the Main Test Set with 1st Trimester Dating Ultrasound (n = 353)

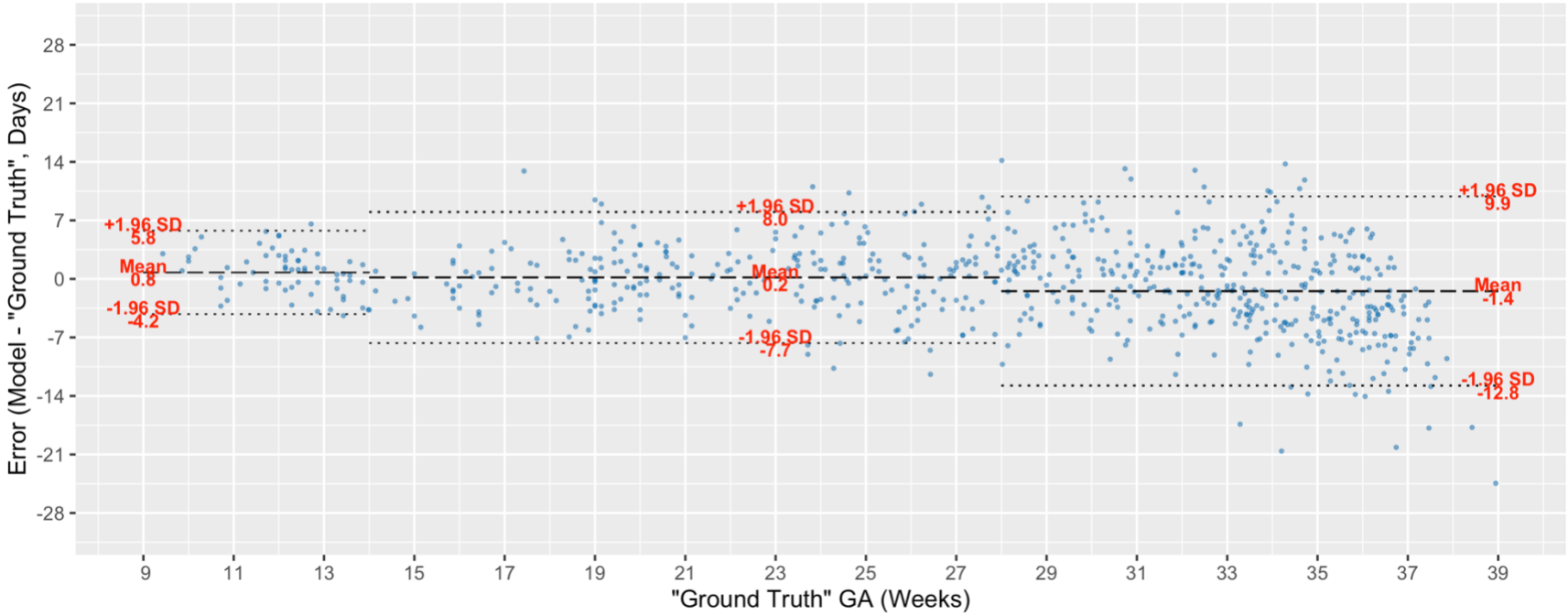


[a] solid line indicates y = x

[b] dashed horizontal lines represent expected error bound of ultrasound biometry according to the American College of Obstetricians and Gynecologists[10] In Zambia, "ground truth" gestational age is defined by the first ultrasound. In North Carolina it is defined by an algorithm incorporating both the last menstrual period and the first ultrasound.[10]

Figure S5: Bland Altman Plot Assessing Agreement between Biometry and Model Gestational Age Estimate in the Main Test Set

Figure S6: Model Error Plot with 95% Limits of Agreement between Model and "Ground Truth" Gestational Age in the Main Test Set



In Zambia, "ground truth" gestational age is defined by the first ultrasound. In North Carolina it is defined by an algorithm incorporating both the last menstrual period and the first ultrasound[10]

Section 4: Supplemental Tables

Table S1: Ultrasound Devices Used

Ultrasound Make and Model	Training Set N=4,770	Tuning Set N=1,188	Main Test Set N=716	IVF Test Set N=47	Novice Test Set N=249
Butterfly iQ (low-cost[a] device)	3833	941	0	0	249
GE LOGIQ C3 Premium	83	17	0	0	0
GE LOGIQ e	136	35	0	0	0
Sonosite MTurbo	1955	457	219	0	0
GE Voluson E8	801	236	188	17	0
GE Voluson S6	1712	417	299	30	0

Each participant study visit involves data collection with both a commercial and low-cost device. We limited the test sets to a single device (Main test set and IVF test set has commercial device only; Novice test set has low-cost device only). We did not impose this limitation on the training and tuning sets (i.e., during training a single participant study could contribute blind sweep cineloops from two devices.) Butterfly = Butterfly Network, Inc Guilford, CT, USA; GE = General Electric Healthcare, Zipf, Austria; Sonosite = SonoSite Inc, Bothell, WA, USA.

[a] the retail price for Butterfly IQ+ device was $2,499 USD in January 2022 (https://store.butterflynetwork.com/us/en/product/butterfly-iq/pro/1-year/; accessed 29 Jan 2022)

Table S2: Gestational Age Estimation of Deep Learning Model Compared to Trained Sonographer in the Novice Test Set

	Novice Test Set (n = 249)[a]				
	Model	Biometry	LMP [b]	Difference Model vs Biometry (95% CI)	Difference Model vs LMP[b] (95% CI)
Mean Absolute Error (SE), days	4.9 (0.29)	5.4 (0.28)	17.4 (1.17)	-0.6 (-1.3, 0.1)	-12.7 (-15.0, -10.3)
Root Mean Square Error, days	6.7 (0.44)	7.0 (0.54)	24.7 (1.68)	-0.3 (-1.7, 1.0)	-18.1 (-21.5, -14.7)
Estimate of Bias [c] (SE), days	-1.4 (0.42)	1.1 (0.44)	-5.5 (1.60)	-	-
1st and 2nd trimester [d,e]					
Mean Absolute Error (SE), days	3.7 (0.39)	4.2 (0.38)	17.2 (2.25)	-0.4 (-1.2, 0.3)	-13.7 (-18.4, -9.1)
Estimate of Bias [c] (SE), days	1.4 (0.57)	2.0 (0.58)	-1.4 (3.14)	-	-
3rd trimester					
Mean Absolute Error (SE), days	5.3 (0.37)	5.9 (0.36)	17.5 (1.37)	-0.6 (-1.6, 0.3)	-12.3 (-15.1, -9.5)
Estimate of Bias [c] (SE), days	-2.5 (0.5)	0.76 (0.57)	-7.0 (1.86)	-	-
Percent Absolute Error < 7 days (SE)	75.1 (2.7)	71.9 (2.8)	40.1 (3.3)	3.2 (-3.3, 9.7)	36.1 (28.0, 44.2)
Percent Absolute Error < 14 days (SE)	95.6 (1.3)	94.8 (1.4)	55.1 (3.3)	0.8 (-2.9, 4.5)	40.5 (33.9, 47.1)

[a] The **novice test set** comprises all participants who contributed at least one set of blind sweeps performed by a novice user on a low-cost, battery-powered device; all participants enrolled in Zambia; expert biometry was performed by a sonographer on a commercial machine.
[b] 22 participants who could not recall their last menstrual period are excluded.
[c] Estimate of Bias is reported as the estimated mean of the error.
[d] Trimesters defined as ≤97 days, 98 – 195 days, ≥196 days.
[e] Only 2 studies in the 1st trimester; 69 studies in the 2nd trimester.
SE=standard error; CI=confidence interval; LMP=last menstrual period

Table S3: Sensitivity Analysis – Gestational Age Estimation of Deep Learning Model Compared to Trained Sonographer in the Main and IVF Test Sets (Allowing Multiple Scans per Participant)

	Main Test Set (n = 1278) [a]			**IVF Test Set (n = 79)** [b]		
	Model	Biometry	Difference (95% CI)	Model	Biometry	Difference (95% CI)
Mean Absolute Error (SE), days	3.9 (0.09)	4.7 (0.11)	-0.8 (-1.1, -0.6)	3.0 (0.26)	3.6 (0.38)	-0.6 (-1.3, 0.1)
Root Mean Square Error, days	5.1 (0.13)	6.2 (0.15)	-1.1 (-1.4, -0.8)	3.8 (0.32)	4.9 (0.47)	-1.1 (-2.0, -0.2)
Estimate of Bias [c] (SE), days	-0.3 (0.14)	1.6 (0.17)	-	0.5 (0.43)	1.4 (0.53)	-
1st trimester [d]						
Mean Absolute Error (SE), days	2.3 (0.17)	2.2 (0.18)	0.1 (-0.4, 0.5)	2.2 (0.34)	2.4 (0.43)	-
Estimate of Bias [c] (SE), days	0.7 (0.26)	1.3 (0.24)	-	0.7 (0.63)	0.4 (0.75)	
2nd trimester [d]						
Mean Absolute Error (SE), days	3.1 (0.11)	3.5 (0.13)	-0.4 (-0.7, -0.2)	2.5 (0.30)	3.0 (0.54)	-
Estimate of Bias [c] (SE), days	0.6 (0.17)	1.4 (0.19)	-	0.9 (0.49)	1.5 (0.69)	
3rd trimester [d]						
Mean Absolute Error (SE), days	4.8 (0.15)	6.1 (0.18)	-1.3 (-1.7, -0.9)	4.2 (0.58)	5.1 (0.76)	-
Estimate of Bias [c] (SE), days	-1.1 (0.23)	1.9 (0.29)	-	-0.3 (1.02)	2.0 (1.21)	
Absolute Error < 7 days (SE), %	85.4 (1.0)	77.3 (1.2)	8.1 (5.6, 10.7)	92.4 (3.0)	84.8 (4.0)	-
Absolute Error < 14 days (SE), %	98.7 (0.3)	96.5 (0.5)	2.3 (1.1, 3.4)	100.0	100.0	-
North Carolina						
Mean Absolute Error (SE), days	3.6 (0.12)	4.1 (0.14)	-0.5 (-0.7, -0.2)	-	-	-
Estimate of Bias [c] (SE), days	-0.1 (0.18)	1.0 (0.20)	-			
Zambia						
Mean Absolute Error (SE), days	4.2 (0.15)	5.5 (0.18)	-1.3 (-1.7, -0.9)	-	-	-
Estimate of Bias [c] (SE), days	-0.4 (0.23)	2.4 (0.28)	-			

Our primary analyses limited test sets to a single ultrasound study per participant. This sensitivity analysis allows participants to contribute more than one study to their test set.

[a] The **main test set** comprises a 30% random sample of participants who are dated by a prior ultrasound and who are not included in the IVF or novice test sets; participants enrolled in either North Carolina or Zambia; blind sweeps and fetal biometry were collected by a sonographer on a commercial ultrasound machine.

[b] The **IVF test set** comprises all studies conducted in women who conceived by *in vitro* fertilization; all participants were enrolled in North Carolina; blind sweeps and fetal biometry were collected by a sonographer on a commercial ultrasound machine.

[c] Estimate of Bias is reported as the estimated mean of the error.

[d] Trimesters defined as ≤97 days, 98 – 195 days, ≥196 days.

SE=standard error; CI=confidence interval

Table S4: Sensitivity Analysis – Gestational Age Estimation of Deep Learning Model Compared to Trained Sonographer in the Novice Test Set (Allowing Multiple Scans per Participant)

	Novice Test Set (n = 330) [a]				
	Model	Biometry	LMP [b]	Difference Model vs Expert (95% CI)	Difference Model vs LMP[b] (95% CI)
Mean Absolute Error (SE), days	5.0 (0.27)	5.5 (0.26)	17.9 (1.06)	-0.5 (-1.1, 0.1)	-13.1 (-15.2, -11.0)
Root Mean Square Error, days	7.0 (0.41)	7.3 (0.49)	25.6 (1.55)	-0.3 (-1.3, 0.8)	-18.8 (-21.9, -15.6)
Estimate of Bias [c] (SE), days	-1.3 (0.38)	1.2 (0.40)	-5.5 (1.45)		
1st and 2nd trimester [d,e]					
Mean Absolute Error (SE), days	3.9 (0.37)	3.9 (0.33)	16.5 (1.90)	-0.0 (-0.7, 0.7)	-13.1 (-17.0, -9.2)
Estimate of Bias [c] (SE), days	1.6 (0.53)	2.1 (0.48)	-0.12 (2.67)		
3rd trimester [d]					
Mean Absolute Error (SE), days	5.4 (0.34)	6.1 (0.33)	18.4 (1.28)	-0.7 (-1.5, 0.1)	-13.1 (-15.6, -10.6)
Estimate of Bias [c] (SE), days	-2.4 (0.46)	0.9 (0.51)	-7.5 (1.71)		
Absolute Error < 7 days (SE), %	74.5 (2.4)	70.9 (2.5)	39.5 (2.8)	3.6 (-2.2, 9.5)	37.2 (30.1, 44.2)
Absolute Error < 14 days (SE), %	93.9 (1.3)	94.2 (1.3)	54.7 (2.9)	-0.3 (-3.7, 3.1)	39.9 (34.0, 45.7)

Our primary analyses limited test sets to a single ultrasound study per participant. This sensitivity analysis allows participants to contribute more than one study to their test set.

[a] The **novice test set** comprises all participants who contributed at least one set of blind sweeps performed by a novice user on a low-cost, battery-powered device; all participants enrolled in Zambia; expert biometry was performed by a sonographer on a commercial machine.

[b] 22 participants (34 studies) who could not recall their last menstrual period are excluded.

[c] Estimate of Bias is reported as the estimated mean of the error.

[d] Trimesters defined as ≤97 days, 98 – 195 days, ≥196 days.

[e] Only 2 studies in the 1st trimester; 88 studies in the 2nd trimester.

SE=standard error; CI=confidence interval; LMP=last menstrual period

Table S5: Sensitivity Analysis – Gestational Age Estimation of Deep Learning Model Compared to Trained Sonographer (Main Test Set Stratified by Gestational Age Basis for "Ground Truth" Pregnancy Dating)

	Main Test Set (n = 716) [a]			
Timing of initial pregnancy dating ultrasound [b]	Model MAE (SE)	Biometry MAE (SE)	Difference (95% CI)	n
1st trimester	3.5 (0.15)	4.0 (0.20)	-0.5 (-0.9, -0.2)	353 [c]
2nd trimester	3.8 (0.17)	5.4 (0.24)	-1.6 (-2.1, -1.0)	279
3rd trimester	6.1 (0.54)	5.3 (0.48)	0.8 (-0.5, 2.0)	84

[a] The **main test set** comprises a 30% random sample of participants who are dated by a prior ultrasound and who are not included in the IVF or novice test sets; participants enrolled in either North Carolina or Zambia; blind sweeps and fetal biometry were collected by a sonographer on a commercial ultrasound machine.

[b] Trimesters defined as ≤97 days, 98 – 195 days, ≥196 days.

[c] n= 322 from North Carolina and n = 31 from Zambia

SE=standard error; CI=confidence interval

Section 5: References for Supplementary Appendix

1. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition; 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
2. Deng J, Dong W, Socher R, Li L, Li K, Li FF. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
3. Kanezaki A, Matsushita Y, Nishida Y. RotationNet for Joint Object Categorization and Unsupervised Pose Estimation from Multi-View Images. *IEEE Trans Pattern Anal Mach Intell.* 2021;43(1):269-283.
4. Su H, Maji S, Kalogerakis E, Learned-Miller EG. Multi-view Convolutional Neural Networks for 3D Shape Recognition. *2015 IEEE International Conference on Computer Vision (ICCV).* 2015:945-953.
5. Ma C, Guo Y, Yang J, An W. Learning Multi-View Representation With LSTM for 3-D Shape Recognition and Retrieval. *IEEE Transactions on Multimedia.* 2019;21:1169-1182.
6. Wang C, Pelillo M, Siddiqi K. Dominant Set Clustering and Pooling for Multi-View 3D Object Recognition. *ArXiv.* 2017;abs/1906.01592.
7. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y. End-to-end attention-based large vocabulary speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016; Shanghai, China; DOI:10.1109/icassp.2016.7472618.
8. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *CoRR.* 2015;abs/1412.6980; available at https://arxiv.org/abs/1412.6980v9 (accessed 23 October 2021).
9. Yao Y, Rosasco L, Caponnetto A. On Early Stopping in Gradient Descent Learning. *Constructive Approximation.* 2007;26(2):289-315.
10. American College of Obstetricians and Gynecologists, American Institute of Ultrasound in Medicine, Society for Maternal-Fetal Medicine. Committee Opinion No 700: Methods for Estimating the Due Date. *Obstetrics and gynecology.* 2017;129(5):e150-e154.