Deep-learning model for prenatal congenital heart disease screening generalizes to community setting and outperforms clinical detection

Athalye, C.; Nisselrooij, A. van; Rizvi, S.; Haak, M.C.; Moon-Grady, A.J.; Arnaout, R.

Ultrasound Obstet Gynecol 2024; **63**: 44–52
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.27503

Deep-learning model for prenatal congenital heart disease screening generalizes to community setting and outperforms clinical detection

C. ATHALYE[1], A. VAN NISSELROOIJ[2], S. RIZVI[1], M. C. HAAK[2], A. J. MOON-GRADY[3] and R. ARNAOUT[1,3,4]

[1]*Division of Cardiology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA;* [2]*Department of Obstetrics, Division of Fetal Medicine, Leiden University Medical Center, Leiden, The Netherlands;* [3]*Department of Pediatrics, Division of Cardiology, University of California, San Francisco, San Francisco, CA, USA;* [4]*Bakar Computational Health Sciences Institute; Department of Radiology; UCSF Berkeley Joint Program in Computational Precision Health; Center for Intelligent Imaging; Biological and Medical Informatics, University of California, San Francisco, San Francisco, CA, USA*

KEYWORDS: artificial intelligence; congenital heart disease; fetal screening; machine learning; ultrasound

CONTRIBUTION

What are the novel findings of this work?
A deep-learning model designed to screen for normal hearts in fetal surveys outperformed experts in a cohort in which over 50% of cases of congenital heart disease (CHD) were initially missed clinically. Notably, the model performed well on community-acquired images in a low-risk population, including lesions on which it had not been trained.

What are the clinical implications of this work?
These findings support the proposition that deep-learning models can improve prenatal detection of CHD.

ABSTRACT

Objectives *Despite nearly universal prenatal ultrasound screening programs, congenital heart defects (CHD) are still missed, which may result in severe morbidity or even death. Deep machine learning (DL) can automate image recognition from ultrasound. The main aim of this study was to assess the performance of a previously developed DL model, trained on images from a tertiary center, using fetal ultrasound images obtained during the second-trimester standard anomaly scan in a low-risk population. A secondary aim was to compare initial screening diagnosis, which made use of live imaging at the point-of-care, with diagnosis by clinicians evaluating only stored images.*

Methods *All pregnancies with isolated severe CHD in the Northwestern region of The Netherlands between 2015 and 2016 with available stored images were evaluated, as well as a sample of normal fetuses' examinations from the same region and time period. We compared the accuracy of the initial clinical diagnosis (made in real time with access to live imaging) with that of the model (which had only stored imaging available) and with the performance of three blinded human experts who had access only to the stored images (like the model). We analyzed performance according to ultrasound study characteristics, such as duration and quality (scored independently by investigators), number of stored images and availability of screening views.*

Results *A total of 42 normal fetuses and 66 cases of isolated CHD at birth were analyzed. Of the abnormal cases, 31 were missed and 35 were detected at the time of the clinical anatomy scan (sensitivity, 53%). Model sensitivity and specificity were 91% and 78%, respectively. Blinded human experts* (n = *3) achieved mean ± SD sensitivity and specificity of 55 ± 10% (range, 47–67%) and 71 ± 13% (range, 57–83%), respectively. There was a statistically significant difference in model correctness according to expert-graded image quality* (P = *0.03). The abnormal cases included 19 lesions that the model had not encountered during its training; the model's performance in these cases (16/19 correct) was not statistically significantly different from that for previously encountered lesions* (P = *0.41).*

Conclusions *A previously trained DL algorithm had higher sensitivity than initial clinical assessment in detecting CHD in a cohort in which over 50% of CHD cases were initially missed clinically. Notably, the DL algorithm performed well on community-acquired images in a low-risk population, including lesions to which it had not been exposed previously. Furthermore, when both the model and blinded human experts had access to only*

Correspondence to: Dr R. Arnaout, 521 Parnassus Avenue Rm 6222, San Francisco, CA 94143, USA (e-mail: rima.arnaout@ucsf.edu)

Accepted: 19 September 2023

stored images and not the full range of images available to a clinician during a live scan, the model outperformed the human experts. Together, these findings support the proposition that use of DL models can improve prenatal detection of CHD. © 2023 International Society of Ultrasound in Obstetrics and Gynecology.

INTRODUCTION

Congenital heart disease (CHD) is the most common birth defect but nevertheless affects only approximately 1% of births per year[1,2]. Prenatal diagnosis of CHD reduces morbidity and mortality and increases therapeutic options. Second-trimester ultrasound examination is recommended universally due to the potential to identify 90% of cases of severe CHD[3,4]. However, in practice, as few as 30% of these are detected[5]. This failure is hypothesized to be due to poor image quality and clinician failure to recognize CHD, and is greater outside of expert centers[5]. Quality improvement programs can increase detection[6–9], but these cannot be applied and sustained universally[5]. Therefore, automated, scalable and robust approaches to prenatal CHD screening are needed.

Previously, we showed that a deep-learning (DL) model could be used to detect CHD[10]. DL, a form of artificial intelligence[11], is a computational method which has the potential for automated and scalable image analysis[12]. Our ensemble model had three steps: view detection on individual images; normal/abnormal decision on individual images by view; and integration of predictions for individual images into a single overall prediction per ultrasound study. This model performed well in two tertiary centers. However, for DL to democratize accurate prenatal detection of CHD[13], it must also perform well in the community. Community imaging may differ from that of tertiary centers: scanning expertise may be lower, and captured images may be fewer and may be stored in low-resolution formats. Different scanning protocols may capture different screening views[4,14] (Figure 1) and may vary among sonographers[15]. Finally, the patient population may vary with respect to several factors, for example the prevalence of CHD and body habitus.

An optimal DL-based fetal screening tool should provide explainability. Our DL model was designed such that the model-learned image features correlated with relevant screening views and important anatomic structures within those views[10], inviting analysis of model performance with respect to these features. However, few community imaging cohorts with these types of annotations exist. In a study by van Nisselrooij *et al.*[16] in 2020, screening ultrasound studies for complex CHD births in 2015–2016 in the northwestern Netherlands were collected and graded for completeness and quality, providing an excellent opportunity to test DL model performance in a well-phenotyped community imaging cohort. We hypothesized that our DL model could be applied successfully to a community-based screening cohort.

METHODS

Datasets

From the PRECOR registry[14,17], ultrasound studies of all pregnancies affected by isolated severe CHD that delivered in the Northwestern region of The Netherlands in 2015 or 2016 were extracted, regardless of whether they had been detected prenatally. Severe CHD was defined by the child requiring surgery in the first year of life. In all cases, the parents were asked to provide consent for collection of images from the second-trimester anomaly scan at the initial community-based screening facility[16]. In order to test the algorithm on images from normal pregnancies as well, we also collected, with consent and from the same screening facilities during the same time period, ultrasound studies without cardiac or other birth defects.

Expert grading

Non-blinded expert grading of ultrasound study quality

Quality assessment for both CHD and normal cases was performed as described previously[16]. Briefly, each of the four cardiac views that was standard in The Netherlands at the time (three-vessel view (3VV), right (RVOT) and left (LVOT) ventricular outflow tract views and four-chamber view (4CV)) was scored for completeness and technical correctness on a scale of 0 to 5 by two fetal echocardiography experts. For each ultrasound study, the experts also noted the duration in minutes and total number of stored images, and graded the fetal position, amount of amniotic fluid, image quality and magnification. Finally, they determined whether the CHD was discernible from the stored images.

Non-blinded annotation of views

To obtain ground-truth labels against which to evaluate the model's performance at the view-sorting step, an expert fetal cardiologist labeled each image frame by view, as described previously[10].

Blinded diagnosis of normal vs *CHD by clinicians based only on stored images*

Three fetal cardiac experts who were blinded to the diagnosis, clinical impression, composition (i.e. proportion that was abnormal) of the dataset and purpose of the study served as the human study subjects. They viewed the stored images and were instructed to grade each heart as normal *vs* CHD. Indicating a specific lesion was optional and was converted to a binary label if provided. Human subjects had access to both color Doppler and grayscale images and videos (Table S1) in their native format (.jpg for still images and .avi for videos). Their fetal cardiac ultrasound experience ranged from 15 years to more than 25 years.

Model inference and screening diagnosis

Clinical images in .jpg (still image) and .avi (video) formats were de-identified and converted to .png, as

described previously[10]. As before, only grayscale images were used (Table S1). Images were input into the ensemble model, which includes a DL view classification step, a DL diagnostic classification step, and a final algorithmic step to integrate model predictions for individual images into an overall ultrasound-study-level decision, as described previously[10] but with the following modifications to the method for filtering low-quality predictions. A neural network's classification of a particular image is actually a set of probabilities of the image belonging to each of the possible classes (these being view classes for the first step of the ensemble, and normal *vs* abnormal classes for the second step[10]); by default, the image is assigned to the class with the highest probability. Previously, only images for which the highest prediction probability was at or above the first quartile of probabilities for that view were deemed of sufficient quality. For the purposes of the current study, entropy across all view predictions per image was calculated[18], such that high entropy indicated model confusion among view categories. High-entropy images were discarded as being of low quality and were not used in the next step of predicting normal *vs* CHD. Similarly, for normal *vs* abnormal prediction for each image, entropy was used to discard low-quality images. An entropy threshold of 0.85 was used, corresponding to the model being at least 70% sure of its diagnostic classification decision. Gradient-weighted class activation maps (GradCAMs) were computed for test images according to standard techniques[19].

Statistical testing

Unless specified otherwise, the Mann–Whitney *U*-test was used for all statistical tests.

Ethics and approval

All investigations were performed in accordance with relevant national guidelines and regulations. All experimental protocols were approved by local institutional review/ethics committees. Participation of clinical experts as human subjects was deemed exempt research by the University of California San Francisco (UCSF) institutional review board (IRB). The ethics board of Leiden University Medical Center (LUMC) approved collection and analysis of images (IRB number P15.374), with written informed consent obtained from all subjects.

RESULTS

Study cohort characteristics

The test dataset included 108 ultrasound studies from patients at 18–22 weeks of gestation, comprising studies of normal hearts and of a range of CHD lesions, as described previously[16] (Table 1). Imaging was collected according to the national protocol in The Netherlands at the time (Figure 1). The mean ± SD total number of items (still images or cine, grayscale and color) per ultrasound study was 41 ± 18 (range, 6–103) and the number was not statistically different between normal and CHD cases (one sample *t*-test, $P = 0.44$). The DL model operates on grayscale images only (Table S1); the number of grayscale still images or cine clips per study was 35 ± 18 (range, 2–78). Therefore, the DL model had fewer stored items on which to base its decision (one-sample *t*-test between all items and grayscale items, $P < 0.01$). Ten CHD ultrasound studies and no normal studies had cine clips stored.

Overall performance

In the test dataset, the sensitivity and specificity of the clinical diagnosis were 53% and 100%, respectively. (Of note, these are not the same as the overall clinical sensitivity and specificity in the northwestern Netherlands[14] due to dataset construction; see Methods.) In contrast, the DL model's sensitivity and specificity on stored grayscale images from this dataset were 91% and 78%, respectively. The model was able to grade 106 of the 108 studies; two studies could not be graded by the model due to insufficient stored image data. Finally, the blinded clinical experts, with access to all stored clinical images, had a mean ± SD sensitivity of 55 ± 10% (range, 47–67%) and specificity of 71 ± 13% (range, 57–83%).

Clinical sensitivity was statistically similar to that of the blinded experts ($P = 0.76$, one-sample *t*-test), while

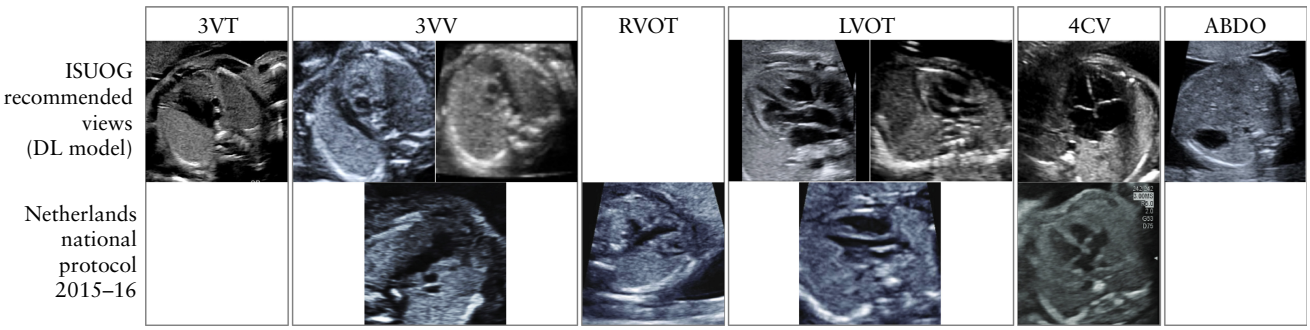

Figure 1 Differences in fetal cardiac view acquisition protocol between development dataset[10] and current study cohort. Three-vessel view (3VV), left ventricular outflow tract (LVOT) view and four-chamber view (4CV) are common to ISUOG and The Netherlands national protocols and represent views used for detection of congenital heart defects in this study. Of note, our deep-learning (DL) model was trained to recognize both 3VV and right ventricular outflow tract (RVOT) view as 3VV and to recognize axial and sagittal LVOT views. 3VT, three-vessel-and-trachea; ABDO, abdomen view. Netherlands national protocol images reprinted from van Nisselrooij *et al.*[16] with permission.

model sensitivity was higher than that of the blinded experts ($P = 0.03$, one-sample t-test). Model specificity was statistically similar to that of the blinded experts ($P = 0.49$, one-sample *t*-test). (Note that because clinically normal studies were specifically chosen in construction of the test dataset, a comparison of specificity between clinical diagnosis and blinded experts is less relevant.)

Performance in CHD cases

In 66 ultrasound studies the fetus had CHD; 35 of these cases were initially detected clinically and 31 were missed. Expert non-blinded retrospective grading determined that, of the 31 misses, 10 were evident on imaging but not recognized, 14 were due to imaging of low technical quality and seven were considered inevitable based on stored imaging, despite adequate imaging quality. Of the 31 CHDs missed clinically, the model detected an anomaly in 27 cases, including five of the seven which were deemed to be inevitable misses, while blinded clinicians detected a mean ± SD of 11 ± 5 (range, 8–17) of the 31 missed cases, including 0–2 of the seven cases considered inevitable. Of 42 clinically confirmed normal fetal hearts, the model identified 32 correctly, while blinded clinicians detected 32 ± 4. These data are summarized in Table 2.

Of the 66 CHD cases, 47 were lesion types that the model had encountered during its training (Table 1), while 19 cases represented lesions that the model had not encountered during training: anomalous left coronary artery from the pulmonary artery, aortopulmonary window, double arch, interrupted aortic arch, pulmonary sling, Shone complex and ventricular septal defect. Despite not having been trained on these lesions, the model detected 16 of these 19 cases.

When the model classifies a particular image, one can visualize the areas in the image most important to the model's decision using GradCAM. For several CHD lesions, the per-image prediction of normal *vs* not normal was largely consistent with clinical knowledge about which views are abnormal in a given lesion (Figures 2 and 3). GradCAMs often, but not always, corresponded to anatomical structures of interest (Figures 2–4). Additionally, we examined failures in model prediction, which can include errors in view

Table 2 Performance of clinical detection, deep-learning model and blinded clinical experts in the Netherlands cohort

Cases	*Clinical diagnosis correct*	*Model decision correct*	*Blinded clinicians correct (mean ± SD)*
Normal ($n = 42$)	42	32	32 ± 4
Abnormal, clinically correct ($n = 35$)	35	32	25 ± 3
Abnormal, clinically missed ($n = 31$)			
Not recognized* ($n = 10$)	0	8	4 ± 2
Technically poor imaging* ($n = 14$)	0	14	6 ± 4
Inevitable*† ($n = 7$)	0	5	1 ± 1
Overall sensitivity (%)	53	91	55 ± 10
Overall specificity (%)	100	78	71 ± 13

Data are given as *n* or % as indicated. *According to non-blinded expert assessment. †Considered inevitable by non-blinded expert graders based on stored imaging and despite adequate imaging quality[16].

Table 1 Congenital heart defects (CHD) present in the Netherlands cohort, indicating lesions included in model training and numbers of lesions detected originally by clinicians as well as by model

	n *(%)*	*Original clinical diagnosis correct* (n)	*Lesion included in model training*	*Model decision correct* (n)
Normal	42 (39)	42	Yes	32
Lesion				
ALCAPA	1 (1)	0	No	0
Aortic stenosis	4 (4)	1	Yes	3
Aortopulmonary window	1 (1)	0	No	1
Atrioventricular septal defect	7 (6)	5	Yes	6
Coarctation of aorta	11 (10)	3	Yes	11
Double outlet right ventricle	4 (4)	4	Yes	4
Double arch	2 (2)	0	No	1
D-TGA	12 (11)	11	Yes	10
Interrupted aortic arch	1 (1)	1	No	1
L-TGA	1 (1)	1	Yes	1
PAIVS	1 (1)	1	Yes	1
Pulmonary sling	1 (1)	0	No	1
Pulmonary stenosis	1 (1)	0	Yes	1
Right atrial isomerism	1 (1)	1	Yes	1
Shone complex	2 (2)	2	No	2
TAPVR	2 (2)	0	Yes	2
Tetralogy of Fallot	2 (2)	1	Yes	2
Truncus arteriosus	1 (1)	1	Yes	1
Ventricular septal defect	11 (10)	3	No	10

ALCAPA, anomalous left coronary artery from pulmonary artery; D-TGA, dextro-transposition of the great arteries; L-TGA, levo-transposition of the great arteries; PAIVS, pulmonary atresia with intact ventricular septum; TAPVR, total anomalous pulmonary venous return.

classification and/or normal/abnormal detection steps (Figure S1); some of these per-image errors did not result in an overall incorrect prediction at the patient level.

Factors affecting ability of clinical examination, model and retrospective blinded human expert review to identify CHD correctly

For detection of CHD clinically, by DL model and by the blinded experts, we tested whether certain ultrasound study features graded and described by van Nisselrooij *et al.*[16] were statistically different from each other based on whether the model and/or clinicians were correct or incorrect. We report *P*-values for these tests in Table 3. For example, study duration in minutes was statistically different by correctness for the initial clinical CHD detection and the DL model but not for the blinded human experts. As in van Nisselrooij *et al.*[16], clinical correctness also varied according to quality and completeness of the cardiac screening views, i.e. with respect to the 4CV and LVOT view quality scores. Overall, quality and completeness of views mattered to blinded human experts as well; for the model, image quality was significant. Number of frames mattered less, both for the model and for the blinded human experts. The performance of only one human expert improved with a greater total number of items. Whether or not non-blinded expert grading considered the diagnosis to be clearly evident in stored images had a statistically significant impact on clinical detection and the blinded human experts, but not the model.

Cine

Though not part of the screening anatomy scan recommendations at the time of the clinical examinations, 10 of the CHD patient studies had cine captures archived. Nine of these fetuses were initially recognized to be abnormal clinically, and the patients with cine stored were statistically more likely to be diagnosed prenatally (9/10 *vs* 27/57, Fisher's exact test $P = 0.016$). The model was correct in all 10 (100%) cases for which cine was available. In the same 10 cases, the human expert reviewers averaged only a 57% pick-up rate, suggesting that cine clips may be only a surrogate for adequate information and do not necessarily contain the information itself.

Model detection of axial screening cardiac views

While for clinical detection and blinded human experts, view detection is implicit, for the DL model view detection is an explicit step. We compared model view detection to that of a non-blinded expert grader as a ground truth. Overall (in both normal and CHD hearts, all grayscale images) the *F*-score comparing model view classification to ground truth was 0.86, representing good agreement. For normal hearts only, the *F*-score was even higher, at 0.96. The *F*-score for CHD hearts only was 0.85. Examples of views detected from both normal and abnormal hearts, along with their corresponding GradCAMs, are shown in Figure 4.

The model is compared against ground truth according to the number of subjects containing a given view, as well as the average number of frames per view, in Table 4.

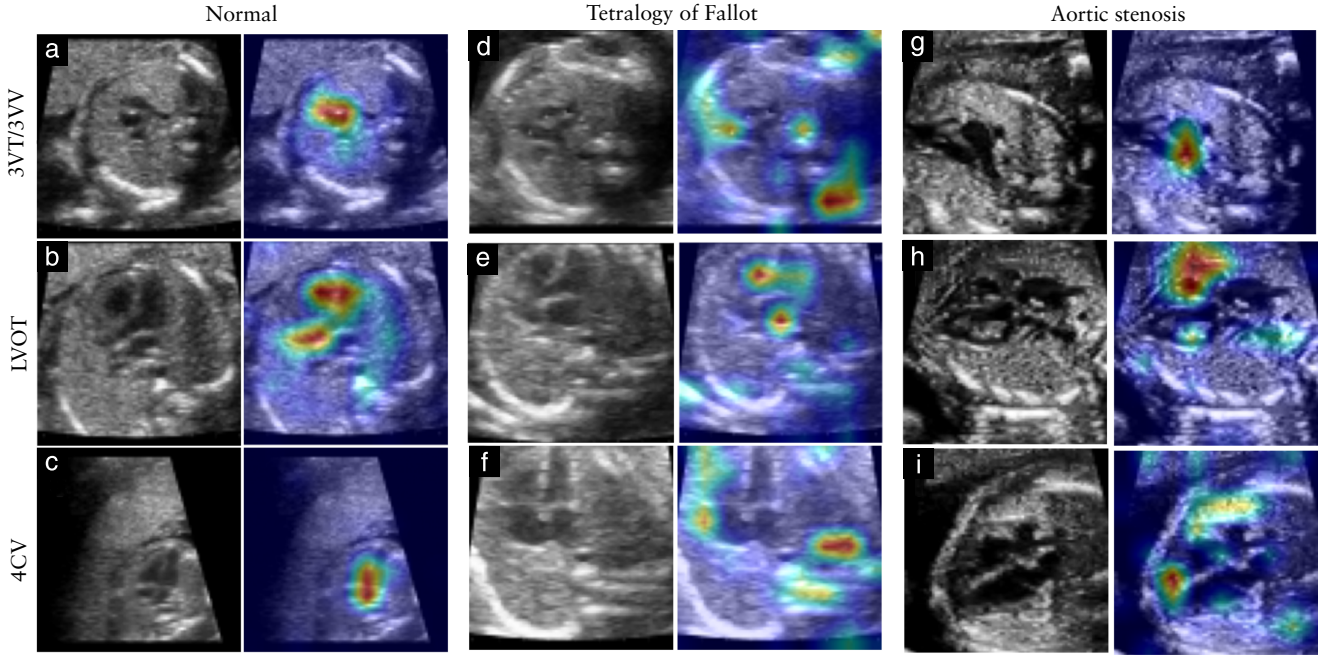

Figure 2 Diagnostic classifier in studies that the model got correct and blinded experts missed. Model-labeled views (grayscale) with corresponding GradCAM images representing heat maps showing areas of the image most important in model decision-making (red shows most important areas). The model correctly identified normal views (a–c), focusing on the aorta (a), left ventricular outflow tract (LVOT) and right ventricle (b), and interatrial and interventricular septa (c). The model identified the abnormal three-vessel-and-trachea (3VT) view and LVOT (d,e) and abnormal four-chamber view (4CV) cardiac axis (f) in tetralogy of Fallot, and abnormal LVOT in aortic stenosis (h). The human experts misclassified these tetralogy and aortic stenosis patients as 'normal'. 3VV, three-vessel view.

Consistent with the *F*-scores above, agreement between model and ground truth was good. In addition to there being more image frames stored for CHD hearts than for normal hearts, there was wide variability as to the number of images per view stored, especially for the CHD hearts, reflecting the presence of cine in some studies.

Finally, although the 3VT view was not part of the imaging protocol for this cohort, 3VT (ground truth) images were present in 14% of the normal studies and 42% of the CHD studies, and these were detected by the model as well (Table 4). Thus, the model can find views even when these are not explicitly acquired per the protocol.

DISCUSSION

Previously, we showed that a DL model could be used to differentiate normal hearts from those with CHD using images from tertiary medical centers[10]. The model design parallels clinician tasks, first finding guideline-recommended views, then classifying images from these views as normal or not normal, and finally aggregating these per-image predictions into a single decision per ultrasound study. In the current work, we expanded testing of the model to anatomy scans obtained in the community, a critical step in ensuring that DL solutions are inclusive of all healthcare settings[13]. These scans were from a group of patients with known outcomes and image-by-image, view-by-view, expert-graded study quality.

Using a community-based cohort that had been well-characterized by non-blinded experts, we compared the model's performance both with clinical detection at the point of care and with the performance of additional human experts who were blinded to the study cohort's composition and outcomes and had access only to stored images. The model had higher sensitivity than did clinical detection, as it flagged the majority of CHD cases that had been missed clinically. Non-blinded expert grading found that the most substantial diagnostic errors arose from either the sonographer's failure to capture adequate (with respect to quality and number) images or the clinician's failure to recognize the abnormality from captured images. Our model represents a potential improvement on clinical performance, being less vulnerable to these obstacles.

While, for the sake of simplicity, the DL model can be said to detect normal hearts *vs* CHD, the model was not, in fact, trained to detect specific CHD lesions (a task already performed quite well by fetal cardiologists)[20,21]. Rather, the model is designed as a screening tool with which to distinguish normal screening ultrasound studies from those that either are abnormal or require further review (e.g. due to incomplete or poor-quality imaging). As such, the model's false-positive rate was high (10/42), and it

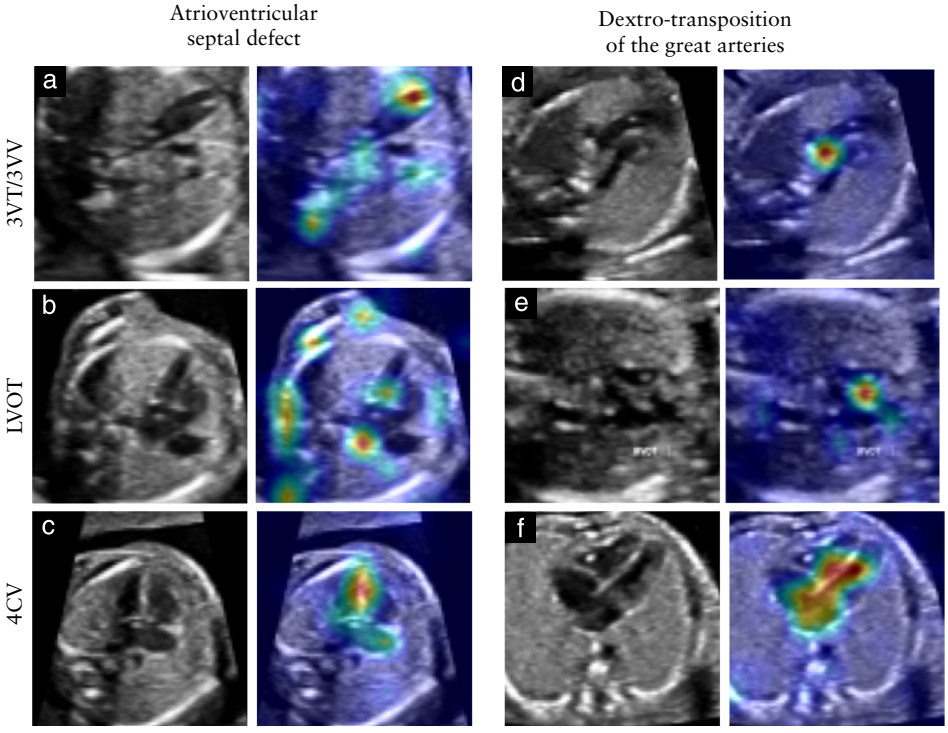

Figure 3 Examples of diagnostic classifier performance in studies that the model got correct but for which blinded expert performance was variable. Model-labeled views (grayscale) with corresponding GradCAM images representing heat maps showing areas of the image most important in model decision-making (red indicates most important areas). (a–c) Atrioventricular septal defect. The model was correct and all three experts recognized the lesion. (d–f) Dextro-transposition of the great arteries (d-TGA). The model identified this study correctly, but two of three experts incorrectly classified it as normal. The GradCAM of the normal-appearing d-TGA four-chamber view (4CV) (f) has similar pattern to normal (Figure 2c), suggesting that the model may function as an anomaly detector. 3VT, three-vessel-and-trachea view; 3VV, three-vessel view; LVOT, left ventricular outflow tract.

cannot currently replace a trained clinician in deciding to refer patients for fetal echocardiography. However, it may be a useful aid for clinicians, decreasing the number of obviously normal ultrasound studies that require review, and flagging studies that are abnormal or that could benefit from more image acquisition at the point of care.

Our study has several strengths. First is the diversity of the imaging cohort[22]. This cohort was external to the dataset on which the model was trained, differed with respect to image formats and scanning protocol (Figure 1) and included imaging studies from several clinics and sonographers. The numbers of still images and cine clips per ultrasound study differed from those in the model's training dataset and were highly variable. Finally, our cohort included a range of CHD lesions, several of which the model had not encountered during training. Despite this diversity, the model's performance was robust and compatible with the model's suspected function as an anomaly detector, which is appropriate for screening. A second strength of our study is the selection of cases from a regional registry which captures CHD cases that have been missed as well as those detected by clinicians. A third strength is our inclusion of blinded human experts to evaluate stored images. While clinical detection was an important comparator, sonographers at the point of care had access to more imaging than that stored. Therefore, the evaluation of blinded clinical experts assessing stored images alone allowed closer comparison to the task that

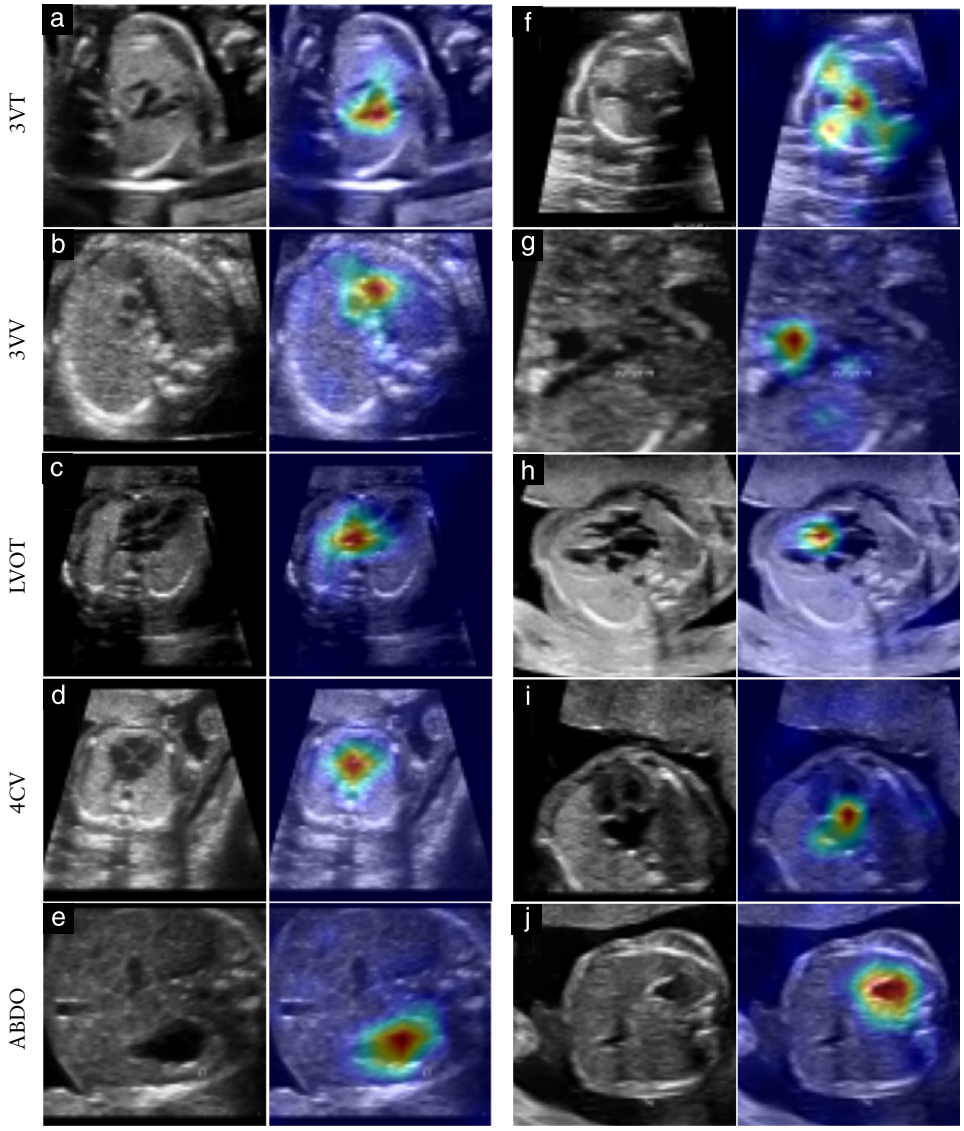

Figure 4 Model view finder is working and clinical features are used in model decisions. Example ultrasound images with corresponding GradCAM images from view-finding step of deep-learning model, illustrating that clinical features are used in model decisions. Images are from fetuses with normal heart (a–e) or congenital heart defect ((f) dextro-transposition of the great arteries (d-TGA); (g) aortic stenosis; (h) d-TGA; (i) levo-TGA; (j) right atrial isomerism). GradCAM images represent heat map showing areas of image most important in model decision-making (red indicates most important areas). (a,f) In both normal and abnormal three-vessel-and-trachea (3VT) views, GradCAM focused on confluence of aortic and ductal arches. (b,g) In normal and abnormal three-vessel views (3VV), GradCAM focused on pulmonary artery and aortic region. (c,h) In normal and abnormal left ventricular outflow tract (LVOT) views, GradCAM focused on LVOT. (d,i) In normal and abnormal four-chamber views (4CV), GradCAM focused on interatrial septum. (e,j) In abdomen view (ABDO), GradCAM focused on stomach.

the model performed. A final strength was our ability to compare clinical, model and blinded human expert performance using a community-based imaging dataset that had been graded for quality and completeness. We found that correctness of the model, like that of the blinded human experts, was associated with quality measures, suggesting that the model's performance was based largely on clinically relevant features. It is interesting that the model performed well in cases of missed CHD that were felt to have been inevitable based on the stored imaging: either the model detected features present in the stored images that were not evident to the human experts, or the model was again acting as an anomaly detector.

Despite its strengths, there were also weaknesses of both the current DL model and the dataset evaluated. Selection of patients for our cohort was limited according to which parents consented to participate. Another important limitation of the model is that it could only evaluate grayscale imaging; in the future, the model may be redesigned to accommodate color imaging. Additional model training and algorithmic improvements may decrease its false-positive rate. However, it is worthwhile noting that the number of model false positives in this study may be at least partly attributable to the limited number of stored images. This theory is supported by the fact that the *F*-score on detected views was high, as well as by the similar specificity of the blinded human experts, who, like the model, lacked additional imaging that would have been available in the clinical setting and were stripped of cognitive bias about the prevalence of CHD in the community. Another limitation of the model might be suggested by the more variable GradCAM results for the model's diagnostic step compared with the view detection step. However, while GradCAM is a useful way to visualize areas of the image that were important to the model in making its decision, and GradCAM heatmaps that focus on anatomical structures of clinical interest is encouraging, a 'poor GradCAM,' i.e. a heatmap that does not focus clearly on clinically relevant features does not necessarily mean that the model's performance is poor. How best to analyze GradCAMs to understand model function is still an active area of research and is beyond the scope of this work[23–25]. Nevertheless, one might

Table 3 *P*-values* for difference between correct and incorrect detection of congenital heart defect (CHD), clinically, by deep-learning (DL) model and by blinded experts, according to ultrasound (US) study characteristics

US study characteristic	*Clinical*	*DL model*	*Blinded expert 1*	*Blinded expert 2*	*Blinded expert 3*
Study duration (in min)	0.01†	0.02†	0.47	0.21	0.38
Number of items (grayscale only, still or cine)	0.44	0.40	0.3	0.36	0.04†
Number of cines	0.08	0.09	0.26	0.5	0.43
Number of cardiac image frames (grayscale only)	0.39	0.46	0.33	0.17	0.03†
Number of 3VV, LVOT and 4CV image frames	0.37	0.42	0.28	0.15	0.03†
Number of 4CV image frames	0.13	0.4	0.29	0.41	0.02†
Diagnosis clear according to non-blinded expert grader	7×10^{-6}†	0.39	< 0.001†	3×10^{-6}†	0.25
View quality score					
Sum of all views	< 0.001†	0.34	0.01†	0.02†	0.05
3VV view quality	0.19	0.15	0.13	0.03†	0.002†
LVOT view quality	< 0.001†	0.44	0.01†	0.03†	0.14
4CV view quality	3.4×10^{-6}†	0.24	< 0.001†	0.17	0.44
Image quality	0.3	0.03†	0.16	0.17	< 0.001†
CHD lesion encountered by model during training	1.4×10^{5}†	0.41	0.04†	0.03†	0.14

*Mann–Whitney *U*-test. †Statistically significant. 3VV, three-vessel view; 4CV, four-chamber view; LVOT, left ventricular outflow tract.

Table 4 Model view detection compared to ground truth

	Ground truth (non-blinded expert)						*Deep-learning model*					
	Subjects containing view			*Frames per subject* (n)			*Subjects containing view*			*Frames per subject* (n)		
View	*Normal*	*CHD*	*Overall*	*Normal*	*CHD*	*Overall*	*Normal*	*CHD*	*Overall*	*Normal*	*CHD*	*Overall*
3VT	6 (14)	28 (42)	34 (31)	0.14 ± 0.35 (0–1)	1.7 ± 6 (0–45)	1 ± 4.7 (0–45)	3 (7)	28 (42)	31 (28)	0.07 ± 0.26 (0–1)	5.4 ± 17 (0–86)	3.3 ± 13 (0–86)
3VV	40 (95)	56 (85)	96 (89)	1.9 ± 1.1 (0–6)	7.6 ± 27 (0–193)	5.4 ± 21 (0–193)	35 (83)	56 (85)	91 (84)	1.4 ± 1.1 (0–5)	10 ± 34 (0–209)	6.9 ± 27 (0–209)
LVOT	33 (79)	56 (85)	89 (82)	1.6 ± 1.5 (0–8)	16 ± 55 (0–304)	11 ± 44 (0–304)	24 (57)	52 (79)	76 (70)	1.2 ± 1.2 (0–5)	37 ± 136 (0–935)	23 ± 108 (0–935)
4CV	41 (98)	64 (97)	105 (97)	2.2 ± 1.7 (0–8)	29 ± 108 (0–807)	19 ± 86 (0–807)	41 (98)	63 (95)	104 (96)	2.2 ± 1.7 (0–2506)	72 ± 327 (0–2506)	45 ± 257 (0–2506)
ABDO	30 (71)	59 (89)	89 (82)	1.1 ± 0.95 (0–4)	1.9 ± 2.7 (0–19)	1.6 ± 2.2 (0–19)	32 (76)	64 (97)	96 (89)	1.8 ± 1.4 (0–5)	11 ± 57 (0–465)	7.3 ± 45 (0–465)
NT	32 (76)	65 (98)	97 (90)	24 ± 17 (0–64)	102 ± 360 (0–2542)	71 ± 284 (0–2542)	34 (81)	66 (100)	100 (93)	23 ± 17 (0–60)	126 ± 372 (0–2535)	85 ± 295 (0–2535)

Data are given as *n* (%) or mean ± SD (range). 3VT, three-vessel-and-trachea view; 3VV, three-vessel view; 4CV, four-chamber view; ABDO, abdomen view; CHD, congenital heart defect; LVOT, left ventricular outflow tract; NT, non-target view.

imagine that a more robustly trained model in the future may yield even better diagnostic performance.

Through this evaluation of stored ultrasound studies, we have demonstrated how variable stored imaging can be, in terms of both numbers of images stored and the views covered. In fact, the model lacked sufficient stored images to analyze two of the studies. While recent recommendations to store cine clips from screening ultrasound examinations are helpful[26], further standardization of stored imaging, through a combination of guidelines and point-of-care integration, will likely improve clinical evaluation as well as computational screening.

In the future, a larger study evaluating a consecutive series of normal and CHD studies in a community population, making use of more standardized image storage and/or integration of a DL model at the point of care, perhaps with an updated DL model, should help to move prenatal screening forward.

ACKNOWLEDGMENTS

C.A., S.R., A.M.G. and R.A. were supported by grants from the National Institutes of Health and the Department of Defense (both grants to R.A.). A.M.G. and R.A. were supported by a generous grant from Georges Harik and Christine Hahn. A.v.N. was supported by a grant from Stichting Hartekind. We thank the clinical experts who served as blinded human research subjects, including Christine Springston, RDCS, and others, who wished to remain anonymous.

REFERENCES

1. CDC. Centers for Disease Control and Prevention. Data and Statistics on Congenital Heart Defects. Published 9 December 2020. https://www.cdc.gov/ncbddd/heartdefects/data.html [Accessed 31 January 2023].
2. Hoffman JIE, Kaplan S. The incidence of congenital heart disease. *J Am Coll Cardiol* 2002; **39**: 1890–1900.
3. Yagel S, Cohen SM, Achiron R. Examination of the fetal heart by five short-axis views: a proposed screening method for comprehensive cardiac evaluation. *Ultrasound Obstet Gynecol* 2001; **17**: 367–369.
4. Carvalho JS, Axt-Fliedner R, Chaoui R, Copel JA, Cuneo BF, Goff D, Gordin Kopylov L, Hecher K, Lee W, Moon-Grady AJ, Mousa HA, Munoz H, Paladini D, Prefumo F, Quarello E, Rychik J, Tutschek B, Wiechec M, Yagel S. ISUOG Practice Guidelines (updated): fetal cardiac screening. *Ultrasound Obstet Gynecol* 2023; **61**: 788–803.
5. Krishnan A, Jacobs MB, Morris SA, Peyvandi S, Bhat AH, Chelliah A, Chiu JS, Cuneo BF, Freire G, Hornberger LK, Howley L, Husain N, Ikemba C, Kavanaugh-McHugh A, Kutty S, Lee C, Lopez KN, McBrien A, Michelfelder EC, Pinto NM, Schwartz R, Stern KWD, Taylor C, Thakur V, Tworetzky W, Wittlieb-Weber C, Woldu K, Donofrio MT, Fetal Heart Society. Impact of Socioeconomic Status, Race and Ethnicity, and Geography on Prenatal Detection of Hypoplastic Left Heart Syndrome and Transposition of the Great Arteries. *Circulation* 2021; **143**: 2049–2060.
6. Brown KL, Sullivan ID. Prenatal detection for major congenital heart disease: a key process measure for congenital heart networks. *Heart Br Card Soc* 2014; **100**: 359–360.
7. Corcoran S, Briggs K, O'Connor H, Mullers S, Monteith C, Donnelly J, Dicker P, Franklin O, Malone FD, Breathnach FM. Prenatal detection of major congenital heart disease - optimising resources to improve outcomes. *Eur J Obstet Gynecol Reprod Biol* 2016; **203**: 260–263.
8. Gardiner HM, Kovacevic A, van der Heijden LB, Pfeiffer PW, Franklin RC, Gibbs JL, Averiss IE, Larovere JM. Prenatal screening for major congenital heart disease: assessing performance by combining national cardiac audit with maternity data. *Heart Br Card Soc* 2014; **100**: 375–382.
9. Letourneau KM, Horne D, Soni RN, McDonald KR, Karlicki FC, Fransoo RR. Advancing Prenatal Detection of Congenital Heart Disease: A Novel Screening Protocol Improves Early Diagnosis of Complex Congenital Heart Disease. *J Ultrasound Med* 2018; **37**: 1073–1079.
10. 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.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–444.
12. Quer G, Arnaout R, Henne M, Arnaout R. Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. *J Am Coll Cardiol* 2021; **77**: 300–313.
13. Morris SA, Lopez KN. Deep learning for detecting congenital heart disease in the fetus. *Nat Med* 2021; **27**: 764–765.
14. van Velzen CL, Clur SA, Rijlaarsdam MEB, Bax CJ, Pajkrt E, Heymans MW, Bekker MN, Hruda J, de Groot CJM, Blom NA, Haak MC. Prenatal detection of congenital heart disease–results of a national screening programme. *BJOG* 2016; **123**: 400–407.
15. Sharma H, Drukker L, Chatelain P, Droste R, Papageorghiou AT, Noble JA. Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. *Med Image Anal* 2021; **69**: 101973.
16. van Nisselrooij AEL, Teunissen AKK, Clur SA, Rozendaal L, Pajkrt E, Linskens IH, Rammeloo L, van Lith JMM, Blom NA, Haak MC. Why are congenital heart defects being missed? *Ultrasound Obstet Gynecol* 2020; **55**: 747–757.
17. Everwijn SMP, van Nisselrooij AEL, Rozendaal L, Clur SAB, Pajkrt E, Hruda J, Linskens IH, van Lith JM, Blom NA, Haak MC. The effect of the introduction of the three-vessel view on the detection rate of transposition of the great arteries and tetralogy of Fallot. *Prenat Diagn* 2018; **38**: 951–957.
18. Xie Q, Luong MT, Hovy E, Le QV. Self-training with Noisy Student improves ImageNet classification. Published online 19 June 2020. https://arxiv.org/abs/1911.04252.
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ArXiv E-Prints*. Published online 1 October 2016. https://arxiv.org/abs/1610.02391.
20. Zhang YF, Zeng XL, Zhao EF, Lu HW. Diagnostic Value of Fetal Echocardiography for Congenital Heart Disease. *Medicine (Baltimore)* 2015; **94**: e1759.
21. van Velzen CL, Clur SA, Rijlaarsdam MEB, Pajkrt E, Bax CJ, Hruda J, de Groot CJM, Blom NA, Haak MC. Prenatal diagnosis of congenital heart defects: accuracy and discrepancies in a multicenter cohort. *Ultrasound Obstet Gynecol* 2016; **47**: 616–622.
22. Chinn E, Arora R, Arnaout R, Arnaout R. ENRICHing medical imaging training sets enables more efficient machine learning. *JAMIA* 2023; **30**: 1079–1090.
23. Martin T. Interpretable Machine Learning. 2019, [Thesis]. https://www.mlmi.eng.cam.ac.uk/files/tam_final_reduced.pdf [Accessed 15 February 2023].
24. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. Published online 6 November 2020. https://arxiv.org/abs/1810.03292.
25. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1**: 206–215.
26. AIUM Practice Parameter for Documentation of an Ultrasound Examination. *J Ultrasound Med* 2020; **39**: E1–E4.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

Figure S1 Poor model performance: errors made by model in view-finding and in diagnostic classification. (a) Correct view (three-vessel view (3VV)) classified incorrectly as abnormal; (b) incorrect view (left ventricular outflow tract (LVOT)) classified appropriately as abnormal; (c) correct view (four-chamber view (4CV)) classified incorrectly as abnormal (possibly due to poor image quality); (d) correct view (three-vessel-and-trachea (3VT) view) classified incorrectly as abnormal (possibly due to low magnification).

Table S1 Imaging data available for clinical decision-making, for blinded human experts and for deep-learning model

