



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** Day, T. G., Budd, S., Tan, J., Matthew, J., Skelton, E., Jowett, V., Lloyd, D., Gomez, A., Hajnal, J. V., Razavi, R., et al (2024). Prenatal diagnosis of hypoplastic left heart syndrome on ultrasound using artificial intelligence: How does performance compare to a current screening programme?. *Prenatal Diagnosis*, 44(6-7), pp. 717-724. doi: 10.1002/pd.6445

This is the published version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/31807/>


**Link to published version:** <https://doi.org/10.1002/pd.6445>

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.



# Prenatal diagnosis of hypoplastic left heart syndrome on ultrasound using artificial intelligence: How does performance compare to a current screening programme?

Thomas G. Day<sup>1,2</sup>  | Samuel Budd<sup>2</sup> | Jeremy Tan<sup>2</sup> | Jacqueline Matthew<sup>2</sup> | Emily Skelton<sup>2,3</sup> | Victoria Jowett<sup>4</sup> | David Lloyd<sup>1,2</sup> | Alberto Gomez<sup>2</sup> | Jo V. Hajnal<sup>2</sup> | Reza Razavi<sup>1,2</sup> | Bernhard Kainz<sup>5,6</sup> | John M. Simpson<sup>1,2</sup>

<sup>1</sup>Department of Congenital Heart Disease, Evelina Children's Healthcare, Guy's and St Thomas' NHS Foundation Trust, London, UK

<sup>2</sup>School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

<sup>3</sup>School of Health Sciences, University of London, London, UK

<sup>4</sup>Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

<sup>5</sup>Department of Computing, Imperial College London, London, UK

<sup>6</sup>Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

## Correspondence

Thomas G. Day, Fetal Cardiology Unit, Ground Floor South Wing, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK.

Email: [Thomas.day@kcl.ac.uk](mailto:Thomas.day@kcl.ac.uk)

## Funding information

National Institute for Health Research, Grant/Award Numbers: NIHR300555, NIHR301448; Wellcome Trust, Grant/Award Number: WT203148/Z/16/Z

## Abstract

**Background:** Artificial intelligence (AI) has the potential to improve prenatal detection of congenital heart disease. We analysed the performance of the current national screening programme in detecting hypoplastic left heart syndrome (HLHS) to compare with our own AI model.

**Methods:** Current screening programme performance was calculated from local and national sources. AI models were trained using four-chamber ultrasound views of the fetal heart, using a ResNet classifier.

**Results:** Estimated current fetal screening programme sensitivity and specificity for HLHS were 94.3% and 99.985%, respectively. Depending on calibration, AI models to detect HLHS were either highly sensitive (sensitivity 100%, specificity 94.0%) or highly specific (sensitivity 93.3%, specificity 100%). Our analysis suggests that our highly sensitive model would generate 45,134 screen positive results for a gain of 14 additional HLHS cases. Our highly specific model would be associated with two fewer detected HLHS cases, and 118 fewer false positives.

**Conclusion:** If used independently, our AI model performance is slightly worse than the performance level of the current screening programme in detecting HLHS, and this performance is likely to deteriorate further when used prospectively. This demonstrates that collaboration between humans and AI will be key for effective future clinical use.

## Key points

### What is already known on this topic?

- Artificial intelligence (AI) can be used to interpret medical images and make diagnoses, including detecting fetal congenital heart disease (CHD) by ultrasound.
- The sensitivity of the current English screening programme for fetal cardiac malformations is publicly available, but specificity is not reported.

Presented preliminary work at British Congenital Cardiac Society Annual Conference 2022, and World Congress of Pediatric Cardiology and Cardiac Surgery 2023.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. Prenatal Diagnosis published by John Wiley & Sons Ltd.

**What this study adds?**

- The current screening programme in our region is operating at a very high specificity for fetal hypoplastic left heart syndrome (HLHS).
- Using a curated retrospective dataset, it is possible to train AI models to detect HLHS with a performance approaching that of the current screening programme.
- Current AI models do not have high enough specificity to be used independently for screening for fetal CHD, meaning that human-AI interaction when performing or interpreting ultrasound will be important to select cases for specialist referral.

## 1 | INTRODUCTION

Congenital heart disease (CHD) is the most common group of congenital malformations.<sup>1-3</sup> Prenatal detection of CHD improves postnatal outcome for some forms of CHD,<sup>4,5</sup> facilitates parental choice, and in selected cases may allow therapeutic intervention in utero.

In the UK, the Fetal Anomaly Screening Programme (FASP) stipulates an offer of an ultrasound scan performed by screening sonographers between 18<sup>+0</sup> and 20<sup>+6</sup> weeks' gestational age (to be completed by 23<sup>+0</sup> weeks), including an examination of the cardiac four-chamber view and outflow tracts.<sup>6</sup> Uptake of the screening ultrasound is 99.1%, indicating excellent access to care.<sup>7</sup> However, current screening programmes fall short of achieving universal detection of CHD. Data from the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS) have recently been published describing the national FASP prenatal detection rates (i.e. the sensitivity of the screening test) in England for a number of congenital anomalies, including some specific CHD lesions.<sup>8</sup> For hypoplastic left heart syndrome (HLHS), the focus of this paper, the current sensitivity is around 92.7% at the anomaly scan, at 94.7% for the screening programme overall.<sup>8</sup>

Artificial intelligence (AI) applied to ultrasound examination of the fetal heart has been explored to improve diagnostic accuracy in the context of fetal CHD, and it has been suggested that some AI models match expert performance levels.<sup>9-11</sup> There is a paucity of information about the impact that AI might have on current CHD screening programmes. To estimate the potential effect of the application of AI in this field, detailed estimates are required of the performance of AI models, and also current national screening programmes.

The aims of our present study were:

- 1) To combine nationally collected data from NCARDRS with data from our own fetal cardiology unit to estimate the current sensitivity and specificity of the FASP in our region in detecting HLHS, as an example CHD lesion.
- 2) To describe the development and performance of an AI algorithm to detect HLHS from ultrasound images.
- 3) To assess how our AI model compares to current screening programme performance, and to consider how such models might be integrated into the current screening system.

## 2 | METHODS

### 2.1 | Estimating performance of the current national screening programme

To assess the current performance of FASP in detecting fetal HLHS, data were combined from NCARDRS and our local clinical unit to calculate true positive, false positive, true negative, and false negative rates. Figure 1 shows a schematic diagram of how these rates are defined. Formulae and figures used to calculate these rates are shown in Supporting Information S1 and Table S1.

True positive and false negative rates were estimated by multiplying the published proportion of HLHS cases that were diagnosed prenatally (including live births, termination of pregnancy, miscarriage, and in utero death) or postnatally, respectively, with an estimate of livebirth prevalence of HLHS from the UK screening committee.<sup>8,12</sup>

To estimate false positive rates, data from our fetal cardiology unit were obtained on the proportion of fetuses that were referred to us following positive FASP screening for CHD that were subsequently found to have a normal heart. This was obtained on a monthly basis for the 6-month period October 2021–March 2022, allowing calculation of a mean with 95% confidence intervals (CIs). This proportion was then multiplied by the previously calculated true positive rate to give a false positive rate.

True negative rates were estimated by summing the true positive, false negative, and false positive rates and subtracting this total from the denominator.

To allow for uncertainty in our estimates, we modelled different scenarios by varying figures used for both the prenatal detection rate of HLHS,<sup>8</sup> and the proportion of fetuses referred to our unit with suspected CHD who were found to have a normal heart. By using the mean value and values at the upper or lower bounds of the 95% CI for these estimates, three estimated scenarios were modelled (i.e. an intermediate scenario, and scenarios that either maximise or minimise current screening performance).

### 2.2 | Artificial intelligence model development

A dataset of fetal echocardiograms was identified by searching our clinical database (Filemaker Pro, Claris Corporation) to identify scans

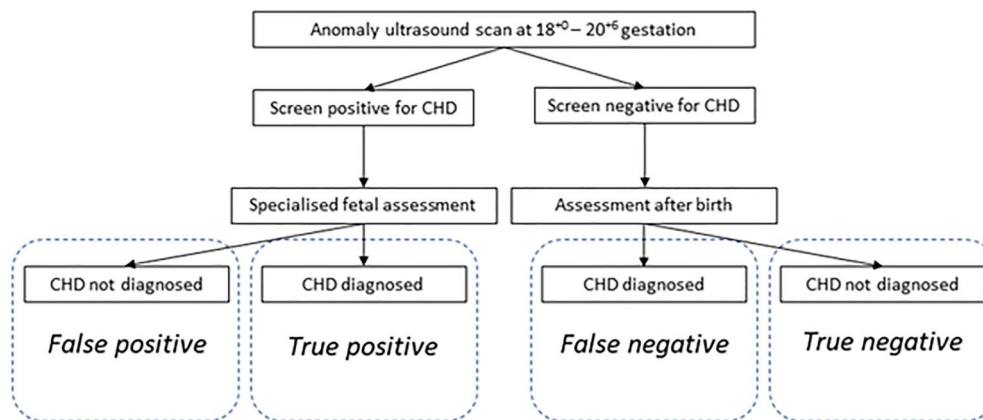


FIGURE 1 Current screening programme for fetal congenital heart disease.

performed between  $20^{+0}$  to  $23^{+6}$  weeks' gestational age, from 01/01/2014 to 31/12/2019. HLHS cases were defined as 'classical HLHS', that is, either mitral atresia with aortic atresia or mitral stenosis with aortic atresia. Fetuses with normal hearts from our database over the same time period and gestation range were also identified, by identifying pregnancies referred for screening due to a family history of CHD, in whom a normal fetal heart was confirmed. Exclusion criteria were multiple pregnancy, or extracardiac anomaly to avoid the AI model inadvertently learning either the extracardiac anomaly appearance, or the normal heart twin which could be labelled as disease. Ultrasound scans were performed using Toshiba Aplio i700/i800 ultrasound machines. Only one scan was used for each fetus.

Demographic and clinical data were extracted for each patient from the same database. Variables were compared between groups using Student's *T* test (for continuous variables) or chi-squared test (for categorical variables).

To identify four chamber (4ch) views, a single expert (TD) identified up to four frames per scan (no more than one per saved clip) that represented an 'ideal' 4ch view according to standard clinical definitions. Using this ideal frame, up to 10 frames chronologically before and after the ideal frame were identified and extracted. Figure 2 shows an example view for a normal fetal heart and HLHS. DICOM imaging data were pseudonymized and converted to Portable Network Graphics (png) image files using a python script.

The image dataset was split on a per-fetus level into training, hyperparameter tuning, and test sets (all images for each fetus were kept in the same set, to ensure there was no contamination of images from the same fetus between sets). Table 1 shows how the dataset was split into different sets. Hyperparameter tuning was performed based on performance on the tuning set. After the final model was selected based on this, for testing 5-fold cross-validation was performed, with the model re-trained for each fold with a different hold-out test set.

Classification models were implemented in Python using Pytorch 1.9 on a PC with a NVIDIA GeForce RTX 3090 GPU. Classification models were based on the ResNet architecture.<sup>13</sup> The final model

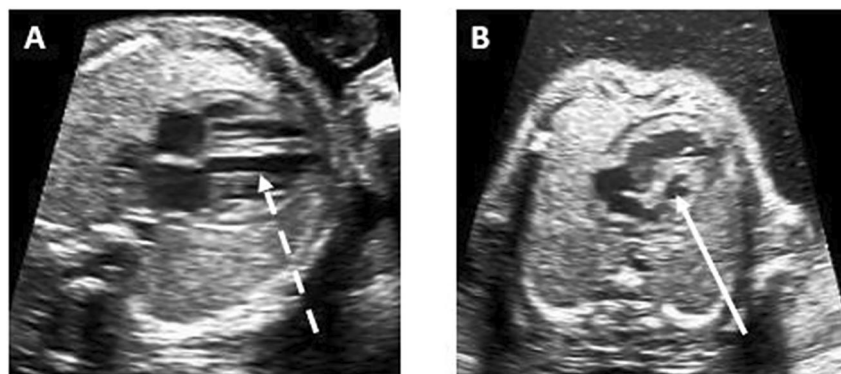
used a ResNet50 architecture, trained with a batch size of 8, for 50 epochs. Batch normalisation was employed. The network was pre-trained using ImageNet. Stochastic gradient descent was used as an optimiser, with a learning rate of 0.001 and momentum of 0.9. A cross entropy loss function was used, weighted to account for the class imbalance in the training set. Images were resampled to  $256 \times 256$  pixels, and normalised. At training time, data augmentation was used (random horizontal flip, random rotation up to  $180^\circ$ , and colour jitter).

Performance of the final model was analysed both on a per-image basis (all images in each test set being classified independently) and on a per-fetus basis, which more closely simulates a real clinical environment (whereby a decision is made on each fetus overall as being affected with CHD or not, using all available images). This was based on a specified proportional threshold of images classified as abnormal, above which the entire fetus is classified as abnormal. This threshold was calibrated, resulting in either a highly sensitive (a lower threshold of abnormal images required to classify the fetus as abnormal), or highly specific (a higher threshold of abnormal images required to classify the fetus as abnormal) model. 95% CIs for sensitivity and specificity were calculated using the exact Clopper-Pearson method for each fold and the mean calculated.

### 3 | RESULTS

#### 3.1 | Performance of the current screening programme in detecting HLHS

Using the intermediate scenario (i.e. using the most likely estimate for each data point to estimate the overall performance), the current screening programme in our region has a sensitivity of 94.3% and specificity of 99.985% in detecting HLHS. Using the maximal performance scenario, it has a sensitivity of 95.9% and specificity of 99.991%, and using the minimal performance scenario, it has a sensitivity of 92.1% and a specificity of 99.976%.



**FIGURE 2** Examples of fetal four-chamber cardiac ultrasound images for (A) normal heart and (B) hypoplastic left heart syndrome. Dashed white arrow: normal left ventricle. Solid white arrow: hypoplastic left ventricle.

**TABLE 1** Numbers of images and fetuses in training, tuning, and testing sets for classification models.

	Total		Normal controls		HLHS cases	
	Fetuses	4ch images	Fetuses	4ch images	Fetuses	4ch images
Training set <sup>a</sup>	130	8260	83	5019	47	3241
Hyperparameter tuning set	15	973	9	593	6	380
Testing set <sup>a</sup>	16	1015	10	676	6	339
Total	161	10,248	102	6288	59	3960

Abbreviations: 4ch, four-chamber view; HLHS, hypoplastic left heart syndrome.

<sup>a</sup>5-fold cross-validation was used during final model testing. The figures here apply to fold 1, and vary slightly for each fold.

### 3.2 | Artificial intelligence to detect HLHS

We trained convolutional neural networks to automatically detect fetal HLHS from a curated retrospective dataset of four-chamber views from fetal ultrasound scans. No significant differences were found between the HLHS and normal control patients for scan gestation, maternal age, maternal ethnicity, or first trimester nuchal translucency (Table 2).

When analysed on a per-image basis, the final models achieved a mean sensitivity of 77.8% (95% CI 73.3%–81.7%) and specificity of 92.7% (95% CI 90.9%–94.7%). When analysed on a per-fetus basis, the highly sensitive model had a sensitivity of 100% (95% CI 50.9%–100%) and specificity of 94% (95% CI 60.4%–99.4%); the highly specific model had a sensitivity of 93.3% (95% CI 49.9%–99.8%) and specificity of 100% (95% CI 68.0%–100%) (Table 3).

Receiver operating characteristic curves were plotted for both the per-image and per-fetus analyses (Figure 3), with a mean area under the curve (AUC) of 0.946 (range 0.917–0.974) for the per-image analysis and a mean AUC of 0.997 (range 0.983–1) for the per-fetus analysis.

### 3.3 | The potential effects of introducing AI to the UK screening programme

Figure 4 and Table 3 show the sensitivity and specificity of the three performance estimates of the current screening programme for HLHS, along with our AI model, calibrated to be either highly

sensitive or highly specific. Extrapolated to the UK, compared to the intermediate estimate of current performance, if operating independently at this performance level our highly sensitive AI model would detect an additional 10 cases of HLHS per year, assuming 756,900 total pregnancies per year (the published 5 years average of 2015–2019 for the UK<sup>14</sup>). However, given the inferior specificity compared to the current system, this would also create an additional 45,134 screen-positive results. The highly specific calibration to the AI model, compared to the intermediate estimate of current performance, would detect two fewer cases of HLHS per year, although with 118 fewer false positive results.

Compared to the minimal performance assumptions of current HLHS screening performance, our highly sensitive AI model would diagnose 14 extra HLHS cases per year, but at the cost of 45,072 more false positive findings. The highly specific calibration AI model would be similar to the minimal real-world performance estimate, with two extra cases of HLHS detected and 180 fewer false positive results (Table 3).

## 4 | DISCUSSION

We have developed an algorithm, utilising deep neural networks, to automate the detection of fetal HLHS. Although our model results are encouraging, comparison of AI performance to current screening programmes is a critical step to assess the potential impact of AI, and to allow consideration of how such models may be incorporated in clinical workflows. This is the first estimate of the current screening

**TABLE 2** Demographic and clinical details for the dataset used for AI model development.

	HLHS cases (n = 59)	Normal controls (n = 102)	p Value
Gestation at time of scan (weeks), mean (SD)	21.0 (1.3)	21.43 (1.3)	0.24
Maternal age (years), mean (SD)	31.3 (5.9)	31.93 (5.8)	0.53
Self-reported maternal ethnicity, n (%)	Any White: 49 (83.1)	Any White: 80 (78.4)	0.37
	Any Black: 2 (3.4)	Any Black: 11 (10.8)	
	Any Asian: 3 (5.1)	Any Asian: 3 (2.9)	
	Other: 1 (1.7)	Other: 1 (1.0)	
	Not recorded: 4 (6.8)	Not recorded: 8 (6.9)	
NT in first trimester (mm), mean (SD)	2.0 (0.9)	1.7 (0.4)	0.06

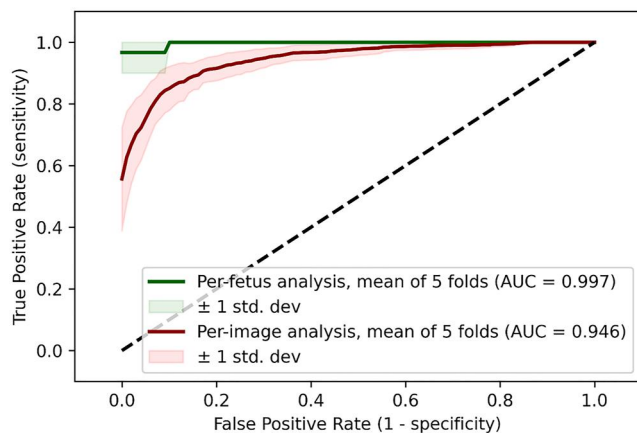
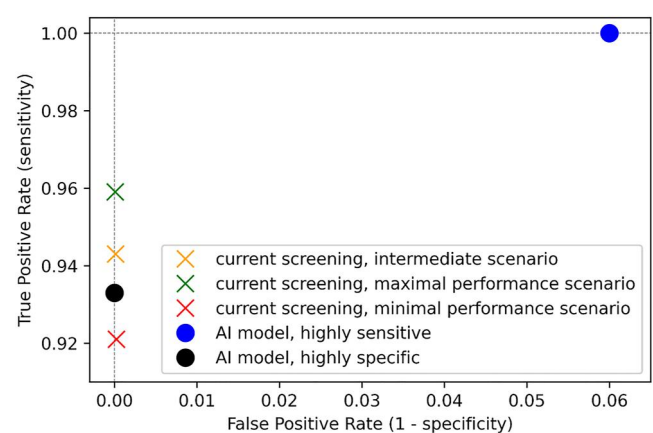
Abbreviations: AI, artificial intelligence; HLHS, hypoplastic left heart syndrome.

**TABLE 3** Performance estimates for the current screening system, and our artificial intelligence models, in detecting fetal hypoplastic left heart syndrome.

							Projected figures per 756,900 pregnancies (UK annual average) <sup>a</sup>			
			Projected figures per million pregnancies <sup>a</sup>							
Model	Sensitivity	Specificity	True positive	True negative	False positive	False negative	True positive	True negative	False positive	False negative
Real-world screening performance estimates										
Intermediate performance scenario	0.943	0.99985	217	999,617	153	13	164	756,610	116	10
Maximal performance scenario	0.959	0.99991	221	999,677	93	9	167	756,656	70	7
Minimal performance scenario	0.921	0.99976	212	999,530	240	18	160	756,544	182	14
Artificial intelligence models										
AI model, calibrated for sensitivity	1.000	0.940	230	940,000	59,770	0	174	711,486	45,240	0
AI model, calibrated for specificity	0.933	1.000	215	999,770	0	15	162	756,726	0	12

Abbreviation: AI, artificial intelligence.

<sup>a</sup>Figures rounded to nearest whole number.

**FIGURE 3** Receiver operating characteristic curves for hypoplastic left heart syndrome detection models, based on per-image analysis and per-fetus analysis. AUC, area under the curve.**FIGURE 4** Estimates of performance of the current UK screening system, and our artificial intelligence models, in detecting hypoplastic left heart syndrome.



programme for a single CHD lesion that includes an estimation of false positives (allowing calculation of screening specificity). The striking feature across all our estimates is the extremely high specificity of the current screening programme for HLHS, meaning only a small number of false positive cases are referred for specialist assessment in fetal cardiology units.

AI has shown promise in the automatic classification of medical images, including fetal CHD on ultrasound.<sup>9</sup> In addition, AI has previously been shown in many contexts to be more reproducible and consistent than human performance.<sup>15</sup> Under current screening programmes, significant variation in the detection rate of CHD has previously been demonstrated across regions, and (although we have not assessed this in the present study) it is possible that AI might become a useful tool to improve consistency, potentially 'levelling up' worse performing areas to match those with higher performance.<sup>16</sup>

If our AI models were used entirely independently of a human diagnostic decision-maker (i.e. AI analysis of human-acquired ultrasound images), compared to the current screening programme the highly sensitive calibration of our AI model would be projected to detect slightly more cases of HLHS per year, but at the cost of many thousands of false positive findings (Table 3). This is unlikely to be compatible with the capacity of specialist services if each false positive resulted in a referral. The highly specific calibration of our model would avoid this problem but may detect slightly fewer cases of HLHS per year than the current screening programme.

However, entirely independent use of such AI models is probably neither desirable nor optimal. As we have discussed in previous work, the real-time, operator-dependent nature of ultrasound means that AI in this context will have to be used in collaboration, with sonographers and AI working together.<sup>17</sup> Used in this way, our AI models could be of benefit, as the human operator acquiring and or interpreting the scan may be able to overrule false positive AI model outputs, but still use the correct AI outputs to improve their sensitivity and reduce false negatives.<sup>18,19</sup>

To achieve this, a local system with human review of images flagged by AI either during or after the scan itself would be required. Although there is evidence that AI may reduce scan time,<sup>20</sup> this might be offset to some degree by additional review time, depending on how AI models were implemented into the clinical workflow. In current practice, screening ultrasound reports are generated at the time of patient attendance, and delay to reporting might also cause additional parental anxiety until a review is taken to pass images as normal or refer on for further assessment.

How the AI model output should best be communicated to the human operator to facilitate this collaboration is still not fully understood, and is an area of active research.<sup>21</sup> As we have previously discussed, trust, and calibration of trust, become very important issues in this context.<sup>17</sup> As shown in the present study, it is likely that any AI model used to detect fetal CHD will be fallible, and will fail in certain situations. For the human-AI team to be effective, the human operator needs to have some way of recognising when this failure has occurred, so they can fall back on their own judgement, but also be appropriately confident that they can trust the AI in most

situations.<sup>22,23</sup> For this to be useful when performing ultrasound, such feedback will likely have to be presented to the human operator in real-time, as the operator may need to modify image acquisition based on the AI model output (taking additional images of a certain cardiac plane, for example). Nevertheless, if a usable and effective feedback system can be developed, then an effective human-AI team might be feasible, even with a less than perfect AI model performance.

However, it is important to consider that the AI model performance presented here (as in other similar work on the topic) is based on a highly curated, retrospectively acquired imaging dataset. In keeping with other fields of medical AI,<sup>24</sup> fetal cardiology suffers from a dearth of research examining how AI might be integrated into real-world clinical practice. There are currently no published studies examining the performance of AI in detecting fetal CHD using prospective data. It is very likely that the performance of our AI models will degrade significantly when translated to the more unpredictable clinical world, where any AI tool will need to be able to detect disease in real-time, based on a stream of ultrasound images obtained by clinicians of varying levels of skill and experience. Prospective clinical trials of any proposed AI model will be essential in order to calculate real-world performance metrics, which will be essential in guiding clinical integration.

The present study has several important limitations. Firstly, because of the rarity of individual CHD lesions, our available retrospective cohort of HLHS fetuses for AI model development is relatively small (59 fetuses). Our AI model performance is similar to previously published models, and we are currently undertaking work to enlarge our imaging dataset, meaning that this problem will be partially ameliorated for future work. However, it will never be possible to match the cohort sizes of hundreds of thousands seen in other branches of medical AI (for example chest x-ray or retinal photograph interpretation<sup>25,26</sup>), so any future methods in fetal cardiology will have to work successfully within this constraint. Our group and others have explored other AI strategies in the context of fetal CHD that may be more resilient to data-scarcity, including image segmentation, normative representation learning, and unsupervised anomaly detection.<sup>27-29</sup> These methods may prove important as they are based on the concept of using only normal cases to train detection models, which may be able to detect abnormal cases despite never being exposed to that particular pathology at training time.

Secondly, the entire imaging dataset used for AI model development was taken from a single centre of fetal cardiology, using only two ultrasound machine models. In addition, our use of training data from a tertiary centre may lead to a model that is not fit for its eventually intended purpose, that is, improving population screening for CHD. This may have implications on the generalisability of this work elsewhere. In addition, we used only a single CHD lesion, HLHS. This was selected as it is clinically severe, but also diagnoseable on a single image plane. HLHS has a relatively gross sonographic appearance, meaning that prenatal detection rates are higher than for some other CHD lesions.<sup>8</sup> However, this likely means that AI detection rates will also be higher, so we feel that comparison between the two methods is



still reasonable. We also used a strict definition of 'classical HLHS' to identify cases for AI training, which may be easier for the AI model to identify compared to other, rarer, subtypes. Further work expanding our findings to other CHD lesions is required.

Thirdly, a major component of our real-world performance modelling was the false positive rate, calculated using our local data. To facilitate referral from sonographers who are not cardiac experts, we do not require referring departments to make a putative diagnosis or explain exactly why they have referred, other than saying they are concerned that the heart is abnormal. This means that these data were not recorded in a disease-specific fashion; hence, we have used a value for CHD as a whole rather than HLHS explicitly. However, given the severity of HLHS, it is unlikely that our estimates of current screening specificity were overly optimistic, as false positive referrals of suspected HLHS are likely to be low.

Fourthly, when estimating current screening performance, we have had access to data that are available at England- or UK-wide level (the prenatal detection rates, and birth prevalence, of HLHS respectively), and data that are only available at a local level (the proportion of screen-positive fetal referrals that were diagnosed with a normal heart). We acknowledge that this means that our performance metrics are very much estimates rather than true measurements, and could best be described as the regional performance of a national screening programme. However, we have tried to account for the uncertainty in our estimates by varying the figures used in the calculations (within reasonable limits), and the main message of extremely high specificity of the current system remains valid.

Finally, as in most work on AI in medical image interpretation, we have only attempted the relatively narrow task of single image analysis. This task does not entirely replicate the cognitive processes performed by a sonographer during live scanning, who will combine non-imaging clinical data, previous imaging examinations, and the complex spatio-temporal relationship between multiple frames in the live video stream to make their final diagnostic decision. We also have not taken into account clinical risk factors that may be of use in improving model performance. We and others are exploring more advanced machine learning methods to harness such additional information, potentially improving AI model performance.

## 5 | CONCLUSION

Our findings suggest that our AI models are probably not good enough to operate independently in diagnosing fetal HLHS on ultrasound images. They may be of benefit in augmenting human performance, but this needs further work to fully assess, particularly in prospective clinical trials.

## ACKNOWLEDGEMENTS

We would like to thank the expert clinicians of the Evelina London Children's Hospital Fetal Cardiology Unit, who performed the ultrasound scans used in this study, and also recorded many of the data points used in the estimates of current UK screening performance.

This work was supported by the Wellcome Trust [IEH Award, 102431], by core funding from the Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z] and by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London and the NIHR Clinical Research Facility. TGD and JM are supported by NIHR Doctoral Fellowships (NIHR301448 and NIHR300555 respectively). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funding bodies had no influence in the study design, data collection, analysis, interpretation, preparation of the manuscript, or decision to submit or publish.

## CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The clinical imaging data used in this study are restricted by ethical data protection permissions and are not available for distribution.

## ETHICS STATEMENT

This study has been granted ethical approval as part of IRAS 149577 REC14/LO/1806. Patients gave written informed consent for the use of anonymised images for research.

## ORCID

Thomas G. Day  <https://orcid.org/0000-0001-8391-7903>

## REFERENCES

- Bernier PL, Stefanescu A, Samoukovic G, Tchervenkov CI. The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.* 2010;13(1):26-34. <https://doi.org/10.1053/j.pcsu.2010.02.005>
- Knowles RL, Bull C, Wren C, Dezateux C. Mortality with congenital heart defects in England and Wales, 1959-2009: exploring technological change through period and birth cohort analysis. *Arch Dis Child.* 2012;97(10):861-865. <https://doi.org/10.1136/archdischild-2012-301662>
- 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(22):2254-2263. <https://doi.org/10.1161/circulationaha.110.947002>
- 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(6):631-638. <https://doi.org/10.1002/uog.14882>
- Bonnet D, Coltri A, Butera G, et al. Detection of transposition of the great arteries in fetuses reduces neonatal morbidity and mortality. *Circulation.* 1999;99(7):916-918. <https://doi.org/10.1161/01.cir.99.7.916>
- NHS Fetal Anomaly Screening Programme Handbook. Public Health England Publications; 2018.

7. Antenatal screening standards: data report 1 April 2019 to 31 March 2020. <https://www.gov.uk/government/statistics/antenatal-screening-standards-data-report-2019-to-2020/antenatal-screening-standards-data-report-1-april-2019-to-31-march-2020>
8. 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*. 2022;130(1):51-58. <https://doi.org/10.1111/1471-0528.17287>
9. 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(5):882-891. <https://doi.org/10.1038/s41591-021-01342-5>
10. Day TG, Kainz B, Hajnal J, Razavi R, Simpson JM. Artificial intelligence, fetal echocardiography, and congenital heart disease. *Prenat Diagn*. 2021;41(6):733-742. <https://doi.org/10.1002/pd.5892>
11. Garcia-Canadilla P, Sanchez-Martinez S, Crispi F, Bijlens B. Machine learning in fetal cardiology: what to expect. *Fetal Diagn Ther*. 2019;47:149-153.
12. Knowles RL, Hunter RM. Newborn Screening for Congenital Heart Defects-External Review against Programme Appraisal Criteria for the UK NSC; 2014.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016:770-778.
14. *Population Estimates for the UK, England and Wales, Scotland and Northern Ireland: Mid-2020*. Office for National Statistics; 2021.
15. Korteling JE, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus artificial intelligence. *Front Artif Intell*. 2021;4:1-13.
16. *National Congenital Heart Disease Audit, Summary Report*. The National Institute for Cardiovascular Outcomes Research; 2021.
17. Day TG, Matthew J, Budd S, et al. Sonographer interaction with artificial intelligence: collaboration or conflict? *Ultrasound Obstet Gynecol*. 2023;62(2):167-174. <https://doi.org/10.1002/uog.26238>
18. Ma S, Lei Y, Wang X, et al. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proc 2023 CHI Conf Hum Factors Comput Syst*. Published online 2023. <https://api.semanticscholar.org/CorpusID:255941863>
19. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. *Int J Hum Comput Stud*. 2023;169:102941. <https://doi.org/10.1016/j.ijhcs.2022.102941>
20. 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*. 2021;42(1):49-59. <https://doi.org/10.1002/pd.6059>
21. 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. Springer International Publishing; 2020:243-252.
22. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1):50-80. <https://doi.org/10.1518/hfes.46.1.50.30392>
23. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res*. 2020;22(6):1-7. <https://doi.org/10.2196/15154>
24. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368:1-12.
25. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J Am Med Assoc*. 2016;316(22):2402-2410. <https://doi.org/10.1001/jama.2016.17216>
26. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Heal*. 2021;3(8):e496-e506. [https://doi.org/10.1016/s2589-7500\(21\)00106-0](https://doi.org/10.1016/s2589-7500(21)00106-0)
27. 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*. Springer International Publishing; 2021:207-217.
28. Chotzoglou E, Day T, Tan J, et al. Learning normal appearance for fetal anomaly screening: application to the unsupervised detection of hypoplastic left heart syndrome. *J Mach Learn Biomed Imag*. 2021;12(September 2021):1-25. <https://doi.org/10.59275/j.melba.2021-g4dg>
29. Tan J, Hou B, Day T, Simpson J, Rueckert D, Kainz B. Detecting outliers with Poisson image interpolation. In: de Bruijne M, Cattin PC, Cotin S, et al., eds. *Medical image computing and computer assisted intervention - MICCAI 2021*. Springer International Publishing; 2021:581-591.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** 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;1-8. <https://doi.org/10.1002/pd.6445>