Radiology: Artificial Intelligence



# A Deep Learning Pipeline for Assessing Ventricular Volumes from a Cardiac Magnetic Resonance Image Registry of Single Ventricle Patients

Tina Yao, MRes*

Nicole St. Clair, BSc*

Gabriel F. Miller, MSc

Adam L. Dorfman, MD

Mark A. Fogel, MD

Sunil Ghelani, MD

Rajesh Krishnamurthy, MD

Christopher Z. Lam, MD

Michael Quail, MD

Joshua D. Robinson, MD

David Schidlow, MD, MMus

Timothy C. Slesnick, MD

Justin Weigand, MD

Jennifer A. Steeden, PhD

Rahul H. Rathod, MD, MBA

Vivek Muthurangu, MD(res)

From the Institutes of Health Informatics (T.Y.) and Cardiovascular Science (M.Q., J.A.S., V.M.), University College London, 20c Guilford Street, London WC1N 1DZ, UK; Department of Cardiology, Boston Children's Hospital, Boston, Mass (N.S.C., G.F.M., S.G., D.S., R.H.R.); Department of Pediatrics, University of Michigan, Ann Arbor, Mich (A.L.D.); Division of Cardiology, The Children's Hospital of Philadelphia, Philadelphia, Pa (M.A.F.); Department of Radiology, Nationwide Children's Hospital, Columbus, Ohio (R.K.); Department of Diagnostic Imaging, Hospital for Sick Children, Toronto, Canada (C.Z.L.); Department of Pediatrics, Ann and Robert H Lurie Children's Hospital of Chicago, Chicago, Ill (J.D.R.); Department of Pediatric Cardiology, Emory University School of Medicine, Atlanta, Ga (T.C.S.); and Department of Cardiology, Texas Children's Hospital, Houston, Tex (J.W.). Received XXX; revision requested XXX; revision received XXX; accepted XXX. Supported by a grant from the Additional Ventures Foundation. T.Y. supported by the UK Research and Innovation Centre for Doctoral Training in AI-enabled Healthcare Systems. **Address correspondence to** V.M. (email: *v.muthurangu@ucl.ac.uk*).

* T.Y. and N.S.C. contributed equally to this work.

https://doi.org/10.1148/ryai.230132

**Purpose:** To develop an end-to-end deep learning (DL) pipeline for automated ventricular segmentation of cardiac MRI data from a multicenter registry of patients with Fontan circulation (FORCE).

**Materials and Methods:** This retrospective study used 250 cardiac MRI examinations (November 2007–December 2022) from 13 institutions for training, validation, and testing. The pipeline contained three DL models: a classifier to identify short-axis cine stacks and two UNet 3+ models for image cropping and segmentation. The automated segmentations were evaluated on the test set ($n$ = 50) using the Dice score. Volumetric and functional metrics derived from DL and ground truth manual segmentations were compared using Bland-Altman and intraclass correlation analysis. The pipeline was further qualitatively evaluated on 475 unseen examinations.

**Results:** There were acceptable limits of agreement (LOA) and minimal biases between the ground truth and DL end-diastolic volume (EDV) (Bias: -0.6 mL/m$^2$, LOA: -20.6–19.5 mL/m$^{2)}$, and end-systolic volume (ESV) (Bias: -1.1 mL/m$^2$, LOA: -18.1–15.9 mL/m$^2$), with high intraclass correlation coefficients (ICC > 0.97) and Dice scores (EDV, 0.91 and ESV, 0.86). There was moderate agreement for ventricular mass (Bias: -1.9 g/m$^2$, LOA: -17.3–13.5 g/m$^2$) and a ICC (0.94). There was also acceptable agreement for stroke volume (Bias:0.6 mL/m$^2$, LOA: -17.2–18.3 mL/m$^2$) and ejection fraction (Bias:0.6%, LOA: -12.2%–13.4%), with high ICCs (> 0.81). The pipeline achieved satisfactory segmentation in 68% of the 475 unseen examinations, while 26% needed minor adjustments, 5% needed major adjustments, and in 0.4%, the cropping model failed.

**Conclusion:** The DL pipeline can provide fast standardized segmentation for patients with single ventricle physiology across multiple centers. This pipeline can be applied to all cardiac MRI examinations in the FORCE registry.



An end-to-end deep learning pipeline was developed to provide automatic segmentation and cardiac function metrics for a cardiac MRI registry of patients with single ventricle physiology. The pipeline requires no human input and is the first to segment this patient population.

### Abbreviations

DL = deep learning, EDV = end-diastolic volume, EF = ejection fraction, ESV = end-systolic volume, SAX = short axis, SV = stroke volume

## Key Points

The developed deep learning segmentation pipeline can provide automated standardized ‘core-laboratory’ processing of a registry of patients with single ventricle physiology that is robust to highly variable anatomy and heterogeneous data collected from > 10 hospitals.

The pipeline achieved median Dice scores of 0.91 (IQR: 0.89–0.94) and 0.86 (IQR: 0.82–0.89) for the end-diastolic and end-systolic blood pool and 0.74 (IQR: 0.70–0.77) for myocardium; there was no evidence of a difference between deep learning and manual measurements for end-diastolic volume, end-systolic volume, myocardial mass, stroke volume and ejection fraction (all $P$ > .05).

The pipeline was further tested on 475 unseen patient examinations, with satisfactory segmentation in both systole and diastole achieved in 68% of examinations, minor adjustments in either systole or diastole needed in 26%, major adjustments needed in 5%, and cropping model failure in only 0.4%.

**Author contributions:**
Guarantor of integrity of entire study, **T.Y., N.S.C., J.D.R., J.W., R.H.R., V.M.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **T.Y., R.K., T.C.S., V.M.**; clinical studies, **N.S.C., A.L.D., M.A.F., R.K., C.Z.L., M.Q., J.D.R., T.C.S., J.W., R.H.R.**; experimental studies, **T.Y., N.S.C., G.F.M., J.D.R., J.A.S., R.H.R.**; statistical analysis, **T.Y., G.F.M., V.M.**; and manuscript editing, **T.Y., N.S.C., A.L.D., M.A.F., R.K., C.Z.L., J.D.R., T.C.S., J.A.S., R.H.R., V.M.**

Conflicts of interest are listed at the end of this article.

Approximately 4–8 in 10000 newborns are born with a functionally single ventricle, and most will undergo a total cavopulmonary connection, leaving them with a Fontan circulation (1). These patients are at risk for cardiac failure, and cardiac MRI is considered the reference standard method of evaluating ventricular size and function. Several small studies have shown that cardiac MRI metrics of ventricular function are predictive of outcomes (2,3), but larger studies are needed to truly understand the importance of cardiac MRI in this patient group (3).

The Fontan Outcomes Registry Using CMR Examinations (FORCE) is the first large-scale (> 4500 cardiac MRI scans in > 3000 unique patients), multicenter cardiac MRI registry of patients who have undergone Fontan palliation (4). Although quantitative ventricular volume data from the original clinical reports are included in the registry, substantial differences in segmentation protocols and interobserver variability make these data unreliable. Unfortunately, standardized ‘core-laboratory’ manual segmentation of the whole registry is neither practical nor feasible, and therefore automated methods are needed to fully harness its potential.

Recent innovations in deep learning (DL) have enabled automated cardiac segmentation methods to reach human levels of accuracy (5–8). However, most DL models are trained and validated on structurally normal hearts (9–11), and automated segmentation in congenital heart disease (CHD) poses a much greater challenge (12,13). Deep learning segmentation models have been successfully developed for biventricular CHD, but none are currently suitable for functionally single ventricles (14,15).

Therefore, we propose a DL pipeline trained on data from the FORCE registry that automatically identifies ventricular short-axis (SAX) cine stacks and crops out the heart and segments the ventricles of patients with Fontan circulation. The pipeline approach is vital for use in registries; recently, Govil et al have demonstrated that such an end-to-end pipeline is feasible

for biventricular CHD (16). Our method extends this approach to complex single ventricular anatomy and applies it to a registry that currently contains data from many sites.

The aims of this study were as follows: i) to develop and validate each section of the pipeline on a curated dataset with "ground truth" segmentations and ii) to test the pipeline on large amounts of registry data, both in terms of segmentation quality and comparison with volumes from the original clinical reports.

## Materials and Methods

This multicenter retrospective study was approved by the institutional review board (IRB) at Boston Children's Hospital, received waiver of consent and was deemed Health Insurance Portability and Accountability Act-compliant (IRB-P00028482). Contributing institutions either relied on the Boston Children's Hospital IRB or gained local IRB/ethics and waiver of consent.

### Pipeline Overview

The automated pipeline (Fig 1) consists of 4 stages: (1) cine stack extraction, (2) SAX identification, (3) heart localization and cropping, and (4) ventricular segmentation and derivation of clinical values. DL models were trained using TensorFlow (v2.12.0) and keras (v2.8.0), and the code is publicly available at https://github.com/Ti-Yao/Single-Ventricle-Segmentation-Pipeline. Hyperband optimization was performed for each model to find which hyperparameters (including network depth, numbers of filters and specific losses) yielded the best validation results (full details in Appendices E2, E3, E4).

### Training Dataset Preparation

The training dataset for all DL models (used in stages 2–4) consisted of complete cardiac MRI examinations of 250 patients from the FORCE registry. This retrospective study was approved by each institution's Committee on Clinical Investigation through a separate application or via reliance agreement, with all examinations de-identified on upload. Patients were scanned at 13 institutions across three countries (USA, UK, Canada) between November 2007 and December 2022, at both 1.5 and 3.0T, using three MRI manufacturers. The dataset was split 175/25/50 for training, validation, and testing. Patients with multiple scans were not split between the training, validation, and testing groups. The training patients were stratified such that the proportion of patients from each site was approximately the same as in the full database (see Appendix E1).

A clinical researcher with 3.5 years of cardiac imaging experience (N.S.) segmented the SAX data to establish ground truth segmentations, which were considered as 'core-laboratory' data. Three cardiovascular imaging physicians (S.G., D.S., R.R.), with 6, 8, and 14 years of experience, respectively, reviewed and adjusted segmentations in 150 cases. Endocardial and epicardial contours at end-systole and end-diastole were manually traced (Circle cvi42 version 5.14.2, Circle Cardiovascular Imaging, Calgary, Alberta, Canada) with trabeculae and papillary muscles included in the blood pool and underdeveloped left or right secondary ventricles contoured. For training of DL networks, traced contours were converted into separate binary masks for the blood pool (combined if two ventricles were present) and myocardium.

### Stage 1: Cine Stack Extraction

Extraction of all cine stacks from each patient study was achieved using Digital Imaging and Communications in Medicine, or DICOM, header information (17). Specifically, two-dimensional (2D) cine data (determined by series with at least 10 frames in the same slice location) were designated to belong to the same stack if cines shared the same orientation (Image Orientation Patient Attribute), pixel size (Pixel Spacing Attribute), and slice thickness (Spacing Between Slices Attribute). Multislice data had to contain at least six slices to be considered a stack.

### Stage 2: Short-axis Identification

A model was trained to select the SAX cine stack from other cine stacks acquired in a single examination (eg, 4-chamber, transverse or long-axis stacks). The training data consisted of the first phase (assumed to be diastolic) of the central five slices of all the cine stacks in the 175 training examinations, labeled either SAX or non-SAX. Nonsquare images were zero-padded to their largest side, and all images were resized to 128 × 128 pixels. The classifier was a convolutional neural network followed by two fully connected dense layers and a sigmoid final layer that outputs the probability that the input image is in the SAX orientation ($P_{sax}$), trained with a binary cross-entropy loss (see Appendix E2 for full details).

At inference, the first phase of the central five slices of each cine stack from each examination was input into the classifier. The stack that contained the slice with the highest $P_{sax}$ was identified as the SAX orientation. If more than one stack had the same maximum probability, then the stack with the highest mean $P_{sax}$ was chosen. For testing (50 examinations), accuracy, precision and recall were assessed per image using a threshold of $P_{sax} > 0.5$ for SAX identification. In addition, the ability to correctly identify the SAX stack per examination was assessed using the process described above.

### Stage 3: Heart Localization and Cropping

Cropping the heart is necessary for a robust pipeline performance, as it centers the heart regardless of its size or position in the original image (18). Cropping the region of interest around the heart was reframed as a simplified segmentation problem. The model was trained to predict ground truth binary ‘whole heart’ masks based on the manually segmented epicardial border. The model was trained using the segmented end-diastolic frames of all slices in the SAX stacks of the 175 training examinations. Nonsquare images were zero-padded to the largest side, and all images were resized to 256 × 256 pixels.

This model was based on a modified UNet 3+ architecture, an improvement on the conventional UNet model that utilizes full-scale skip connections and can use deep supervision (19). The UNet 3+ was trained using Intersection over Union (IoU) loss with deep supervision (the model parameters were determined by Hyperband optimization; see Appendix E3 for full details).

The steps for using the segmentation to crop images at inference are shown in Figure 2. In step 1, the model predicts the ‘whole heart’ masks in the first (assumed to be diastolic) phase

for all slices in the SAX stack. In step 2, disconnected 'islands' in the predicted masks (misclassified regions such as the stomach) were removed if they did not overlap with the intersection of the masks in the slice direction. In step 3, a preliminary bounding box was defined as the minimum square containing the union of all the segmentation masks, which was then expanded by 50% for redundancy. In step 4, the bounding box was used to crop all slices and frames in the stack. Ground truth bounding boxes were derived from ground truth 'whole heart' masks as described above and were compared with the predicted bounding boxes in the 50 test cases using IoU.

### Stage 4: Ventricle Segmentation

The final segmentation model also used the UNet 3+ architecture and was trained to predict three pixel classes: blood pool, myocardium and background. The model was trained on all the end-diastolic and end-systolic slices in the SAX stacks of the 175 training examinations (4308 images). Images that did not contain ventricular structures were labeled as all background. All images were cropped using bounding boxes created from the ground truth masks as described above and resized to 128 × 128 pixels. The UNet 3+ was trained using Tversky loss, but without deep supervision (the model parameters were determined by Hyperband optimization; see Appendix E4 for full details).

At inference, all slices and frames of the SAX stack were segmented. Postprocessing included removing any component separate from the heart, where the heart is defined as the largest connected component in 3D. From these masks, ventricular volume-time curves were created by summing the labeled blood pool voxels (scaled by voxel volume) across all slices for all cardiac phases. End systole and diastole were identified as the phases with the smallest and largest volumes using the volume time curve, constrained to the middle five slices (to lessen any effect of poor segmentation at the base or apex). End-systolic volume (ESV) and end-diastolic volume (EDV) were then defined as the volume (assessed over all slices) at these timepoints. Stroke volume (SV) was EDV–ESV, and ejection fraction (EF) was SV/EDV. Myocardial mass was the sum of the myocardial voxels (scaled by voxel volume and myocardial density) in the end-diastolic phase. All volumes and mass were indexed to body surface area (BSA).

For testing (50 examinations), the DL-predicted and ground truth blood pool and myocardial masks were compared per slice using Dice at end-diastole and end-systole. DL-derived ESV, EDV, SV and EF were compared with the results from manual core-laboratory segmentation using Bland Altman and correlation analysis. Additionally, we calculated the sensitivity and specificity of identifying a dilated ventricle (defined as BSA-indexed EDV ≥ 156 mL/BSA$^{1.3}$ (3)) using DL volumes, with the volumes derived from manual segmentation serving as the ground truth.

### Pipeline Performance

The performance of the whole pipeline was tested by processing 475 unseen examinations (not used for previous training or testing)-including data from six sites not represented in the training data (site breakdown in Appendix E1). All end-diastolic and end-systolic segmentations were examined by the clinical researcher, and the pipeline results were rated as follows: i) satisfactory

segmentation (appropriate for use in a clinical context), ii) segmentation requiring minor adjustments (small adjustment required in 1–2 slices), iii) segmentation requiring major adjustment (significant failure in > 2 slices), and iv) crop failure (part or all of the heart missing from cropped images). In addition, subjective image quality was rated as satisfactory or suboptimal depending on the presence of artifacts, poor contrast, or very high noise See Appendix E5 for examples of segmentation and subjective image quality ratings. In those studies deemed to have satisfactory segmentation or only requiring minor adjustment, DL-derived volumetric data were compared with the clinical report data entered into the FORCE registry by each contributing site. Clinical report data were also compared with the manual core-laboratory-derived volumes and EF in the training dataset to evaluate reliability.

### Statistical Analysis

Continuous variables are expressed as median (interquartile range-IQR), as most variables were not normally distributed. Bland-Altman and intraclass correlation (reported with 95% confidence intervals-CI) were used to assess agreement between DL, manual core-laboratory and clinical report volumetric data. As the differences between the manual and DL measurements were normally distributed (evaluated using Shapiro-Wilk test), paired *t* tests were used to assess significance. Comparison of pipeline results for different ventricular morphologies, pediatric versus adult, magnetic field strengths (1.5T versus 3T), dilated versus smaller ventricles (defined as above (3)) and scans periods (pre-2015 versus post-2015, where 2015 is the midpoint of the scan time range), was performed using the $\chi^2$ test. Statistical analyses were performed using the scipy library (v1.9.1) in Python and in R, and $P < .05$ was considered statistically significant.

## Results

### Dataset Characteristics

Training of models was successfully completed for all stages of the pipeline. Table 1 presents the demographics for the 250 (median [IQR] age: 16 [11–22], 154 male, 97 female) patients in the training and validation sets and the 475 patients (median [IQR] age: 14 [9–19], 276 male, 202 female) in the external test set. There was no evidence of differences in age, BSA or ventricular morphology between the training, validation, or test datasets (see Appendix E6).

### Short-axis Identification Classifier and Heart Localization Model

The accuracy for SAX identification per slice was 96.1%, precision was 98.0%, and recall was 94.4%. However, because all slices are evaluated to make a final decision, the classifier was able to correctly identify the short-axis stack in all 50 test examinations.

For heart localization, the median IoU between the ground truth and predicted bounding boxes per examination was 0.94 (IQR: 0.92–0.96). More importantly, the calculated bounding box contained the whole heart for all 50 test examinations.

### Ventricle Segmentation Model

Figure 3 shows examples of DL ventricular segmentation for the best, median and worst cases (regardless of systolic or diastolic phase) based on Dice score (see Appendix E7 for movies of all frames).

There were acceptable limits of agreement and high levels of intraclass correlation between DL and manual core-laboratory segmented EDV and ESV (Table 2, Fig 4). EDV and ESV were slightly higher with DL segmentation, but the biases were not clinically important (-0.6 mL/m$^2$ and-1.1 mL/m$^2$ respectively) and did not reach statistical significance ($P \geq$ 0.56). There was also acceptable agreement for SV and EF and moderate agreement for ventricular mass (Table 2, Fig 5), with small biases that did not reach statistical significance ($P \geq .12$). The sensitivity and specificity for detecting a dilated ventricle was 94% (137/145) and 79% (262/330) respectively. The median Dice scores for the blood pool (diastolic and systolic) and myocardial masks were 0.91 (IQR: 0.89–0.94), 0.86 (IQR: 0.82–0.89) and 0.74 (IQR: 0.70–0.77) respectively (Table 2).

### Pipeline Performance

Pipeline processing was feasible in all 475 new examinations, with all SAX stacks correctly identified. Table 3 shows the time taken and the number of images processed at each stage. The average time to process a patient examination through the pipeline was 26s (range: 21–32s).

For all examinations, end-diastolic and end-systolic frames were successfully extracted, totaling 950 volumes. Of these, 767 frames (81%) met the criteria for satisfactory segmentation quality. At the examination level, out of a total of 475 examinations, 323 (68%) exhibited satisfactory segmentation for both end-systolic and end-diastolic volumes. Minor adjustments were required for 124 (26%) examinations, major adjustments for 26 (5%) examinations, and the cropping model failed in 2 (0.4%) examinations (see Appendix E5 for examples).

Approximately 35% of SAX stacks were identified as having suboptimal image quality, and there was a statistically lower ($P < .001$) proportion of satisfactory segmentations in this group (Table 4). Segmentation was also more successful in 1.5T images compared with 3.0T ($P < .001$). We found no evidence of a difference between segmentation success for different ventricular morphologies ($P = .11$), different vendors ($P = .13$), pediatric versus adult patients ($P = .36$), dilated versus smaller ventricles ($P = .56$) or pre-2015 versus post-2015 ($P = .13$).

In the 475 patients, there was only moderate agreement between DL-generated volumes and EF and clinical report data (Fig 6), with DL volumes being significantly higher ($P < .001$), particularly at higher volumes. It should be noted that in the training data ($n = 250$), we observed similar results for the levels of agreement and biases ($P < .001$) between manual core-laboratory and clinical report volumes and EF (Fig 6).

## Discussion

To our knowledge, this study is the first description of DL automated segmentation of single ventricles. The main findings were as follows: i) It is feasible to create an end-to-end deep learning pipeline that automatically takes cardiac MRI examinations, extracts SAX cines, performs cropping, and segments the ventricles; ii) There was acceptable agreement between DL

and manual segmentation of functional single ventricles in terms of volumes and mass; iii) The pipeline can rapidly and accurately segment large numbers of unseen cases with a high degree of success; and iv) There was only moderate agreement between volumetric data from clinical reports and both DL and manually segmented core-laboratory data, implying that data from clinical reports is highly variable. Our framework processes images straight from the image registry (16) and could process the whole FORCE registry (> 4500 examinations) in < 40 hours.

Although the use of DL for ventricular segmentation in cardiac MRI is well described, its use in CHD is much more limited, mainly due to less access to large training datasets and more complex anatomy (13). Nevertheless, successful models have been produced (12,14–16) for biventricular CHD, including pipeline frameworks for Tetralogy of Fallot (16). However, creating a pipeline that is capable of segmenting extremely heterogeneous single ventricular anatomy from a large number of institutions with a wide range of protocols, scanners, and field strengths is substantially more difficult. The size of the FORCE registry allowed for creation of a large training dataset collected from 13 different hospitals, aiding generalizability. Furthermore, we choose to use the UNet 3+ architecture, which has been shown to improve segmentation accuracy over simpler UNet architectures.

Our DL segmentation pipeline demonstrated acceptable agreement with manual segmentation for ventricular volumes, mass, and EF, with results comparable to other DL models for cardiac MRI segmentation in patients with CHD (12,14–16). Importantly, DL always produces the same result, which is highly desirable, as it has been shown that there can be significant interobserver variability of cardiac MRI metrics in patients with Fontan circulation (20).

When applied to 475 unseen cases from 16 institutions (including data from 6 sites not included in training), our end-to-end pipeline was successful in 68% of cases, with only a minor proportion requiring full resegmentation. This means that if our pipeline was applied across the whole FORCE registry, > 4500 examinations could be robustly processed without any user input. Unsurprisingly, DL performance was lower in studies with poor image quality, but human segmentation would also be affected in these cases. In addition, performance was slightly lower at 3T, likely due to different signal characteristics, increased artifact, and the lower proportion of 3T studies in the training data.

Further improvements could be achieved by fine-tuning the UNet 3+ with manually segmented images from studies where the DL model has failed (including poor image quality and 3T studies). Such an approach would benefit from automated quality assurance, which could determine segmentation accuracy (21,22) and automatically identify data that require resegmentation.

It should be noted that the agreement between the DL-derived and clinical report volumes was only moderate, with DL volumes being higher than those entered by the site. However, these differences were also present between manually segmented core-laboratory data and the clinical report volumes. This suggests that differences in segmentation between sites (eg, inclusion of papillary muscle in blood pool and segmentation of underdeveloped ventricle), as well as interobserver variability, were the main causes of disagreement.

The main limitation of this study was that manual segmentations and qualitative assessment were performed by a single operator. This was done to ensure consistency of segmentation and assessment but could result in biases in the models. However, we believe that the benefit of having a single adequately trained observer (with expert review when necessary) is that the DL model is trained on highly homogeneous data and results are evaluated in a consistent manner. A further limitation of our study is the lack of normal biventricular controls. However, it has been shown previously that models trained on congenital biventricular data struggle to robustly segment single ventricle data (14). Consequently, we expect the opposite to be true with our model.

To conclude, we have demonstrated a pipeline for automated segmentation of ventricular volumes from cardiac MRI scans of patients with Fontan circulation. We believe that combining the FORCE registry's clinical data with these automatically derived cardiac metrics will give researchers and clinicians new insights into the role of cardiac MRI in the management of patients with Fontan circulation. Future work will include developing DL models to evaluate strain, which has also been found to be predictive of outcome (3).

**Disclosures of conflicts of interest: T.Y.** PhD studentship is funded by UK Research and Innovation (UKRI), which is the UK government research funding body. **N.S.C.** No relevant relationships. **G.F.M.** No relevant relationships. **A.L.D.** Grant from Additional Ventures; Additional Ventures via the FORCE grant, support for attending the FORCE Executive Board meeting at the SCMR Scientific Sessions 2023. **M.A.F.** Grant support - 5% effort from Additional Ventures Foundation; NIH Roi grant; grant from Additional Ventures for another project (grant on single ventricles but nothing to do with the current paper); grant from Rocket Pharma (for MRI core lab for Danon disease); royalties from book sales from Wiley; expert witness for law firm; CMP Pharma contributes drug for a single ventricle study. **S.G.** No relevant relationships. **R.K.** No relevant relationships. **C.Z.L.** The FORCE registry receives grant funding from Additional Ventures; member of the FORCE registry Executive Leadership Team. **M.Q.** No relevant relationships. **J.D.R.** Salary support from Fontan Outcomes Registry Using CMR Examinations (FORCE) which is supported by a grant from Additional Ventures. **D.S.** No relevant relationships. **T.C.S.** No relevant relationships. **J.W.** No relevant relationships. **J.A.S.** Funding from UKRI: Future Leaders Fellowship. Grant Ref: MR/S032290/1; Educational Board Member for ISMRM (International Society for Magnetic Resonance in Medicine); Committee member for EACVI; Exam board representative in for

pediatric EuroCMR; Chair of Artificial Intelligence (AI) Subcommittee for Paediatric SCMR. **R.H.R.** This research was supported through a grant by the Additional Ventures Foundation. **V.M.** Grant funding to university from British Heart Foundation; Exam board of the European Association of Cardiovascular Imaging Congenital and Paediatric CMR accreditation.

## References

<jrn>1. O'Leary PW. Prevalence, clinical presentation and natural history of patients with single ventricle. Prog Pediatr Cardiol 2002;16(1):31–38.</jrn>

<jrn>2. Poh CL, d'Udekem Y. Life After Surviving Fontan Surgery: A Meta-Analysis of the Incidence and Predictors of Late Death. Heart Lung Circ 2018;27(5):552–559.</jrn>

<jrn>3. Meyer SL, St Clair N, Powell AJ, Geva T, Rathod RH. Integrated Clinical and Magnetic Resonance Imaging Assessments Late After Fontan Operation. J Am Coll Cardiol 2021;77(20):2480–2489.</jrn>

<eref>4. Fontan | Force Study. Fontan Outcome Registry Using CMR Examinations. FORCE Website. https://www.forceregistry.org. Accessed February 14, 2023.</eref>

<jrn>5. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018;20(1):65.</jrn>

<edb>6. Isensee F, Jaeger PF, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features. In: Pop M, Sermesant M, Jodoin PM, et al, eds. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. STACOM 2017. Lecture Notes in Computer Science, vol 10663. Springer, 2018; 120–129.</edb>

<jrn>7. Duan J, Bello G, Schlemper J, et al. Automatic 3D Bi-Ventricular Segmentation of Cardiac Images by a Shape-Refined Multi- Task Deep Learning Approach. IEEE Trans Med Imaging 2019;38(9):2151–2164.</jrn>

<jrn>8. Backhaus SJ, Staab W, Steinmetz M, et al. Fully automated quantification of biventricular volumes and function in cardiovascular magnetic resonance: applicability to clinical routine settings. J Cardiovasc Magn Reson 2019;21(1):24.</jrn>

<jrn>9. Petitjean C, Zuluaga MA, Bai W, et al. Right ventricle segmentation from cardiac MRI: a collation study. Med Image Anal 2015;19(1):187–202.</jrn>

<eref>10. Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI. The MIDAS Journal. . Published July 9, 2009. Accessed February 14, 2023.</eref>

<jrn>11. Bernard O, Lalande A, Zotti C, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Trans Med Imaging 2018;37(11):2514–2525.</jrn>

<edb>12. Wolterink JM, Leiner T, Viergever MA, Išgum I. Dilated Convolutional Neural Networks for Cardiovascular MR Segmentation in Congenital Heart Disease. In: Zuluaga M, Bhatia K, Kainz B, Moghari M, Pace, D, eds. Reconstruction, Segmentation, and Analysis of Medical Images. RAMBO HVSMR 2016 2016. Lecture Notes in Computer Science, vol 10129. Springer, 2017; 95–102.</edb>

<jrn>13. Van den Eynde J, Kutty S, Danford DA, Manlhiot C. Artificial intelligence in pediatric cardiology: taking baby steps in the big world of data. Curr Opin Cardiol 2022;37(1):130–136.</jrn>

<jrn>14. Karimi-Bidhendi S, Arafati A, Cheng AL, Wu Y, Kheradvar A, Jafarkhani H. Fully-automated deep-learning segmentation of pediatric cardiovascular magnetic resonance of patients with complex congenital heart diseases. J Cardiovasc Magn Reson 2020;22(1):80.</jrn>

<jrn>15. Tandon A, Mohan N, Jensen C, et al. Retraining Convolutional Neural Networks for Specialized Cardiovascular Imaging Tasks: Lessons from Tetralogy of Fallot. Pediatr Cardiol 2021;42(3):578–589.</jrn>

<jrn>16. Govil S, Crabb BT, Deng Y, et al. A deep learning approach for fully automated cardiac shape modeling in tetralogy of Fallot. J Cardiovasc Magn Reson 2023;25(1):15.</jrn>

<jrn>17. Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc 1997;4(3):199–212.</jrn>

<jrn>18. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D Consistent and Robust Segmentation of Cardiac Images by Deep Learning With Spatial Propagation. IEEE Trans Med Imaging 2018;37(9):2137–2148.</jrn>

<prpt>19. Huang H, Lin L, Tong R, et al. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. arXiv 2004.08790 [preprint] https://arxiv.org/abs/2004.08790. Published April 19, 2020. Accessed February 14, 2023.</prpt>

<jrn>20. Margossian R, Schwartz ML, Prakash A, et al. Comparison of echocardiographic and cardiac magnetic resonance imaging measurements of functional single ventricular volumes, mass, and ejection fraction (from the Pediatric Heart Network Fontan Cross-Sectional Study). Am J Cardiol 2009;104(3):419–428.</jrn>

<jrn>21. Bartoli A, Fournel J, Bentatou Z, et al. Deep Learning-based Automated Segmentation of Left Ventricular Trabeculations and Myocardium on Cardiac MR Images: A Feasibility Study. Radiol Artif Intell 2020;3(1):e200021.</jrn>

<edb>22. Robinson R, Oktay O, Bai W, et al. Real-Time Prediction of Segmentation Quality. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11073. Springer, 2018; 578–585.</edb>

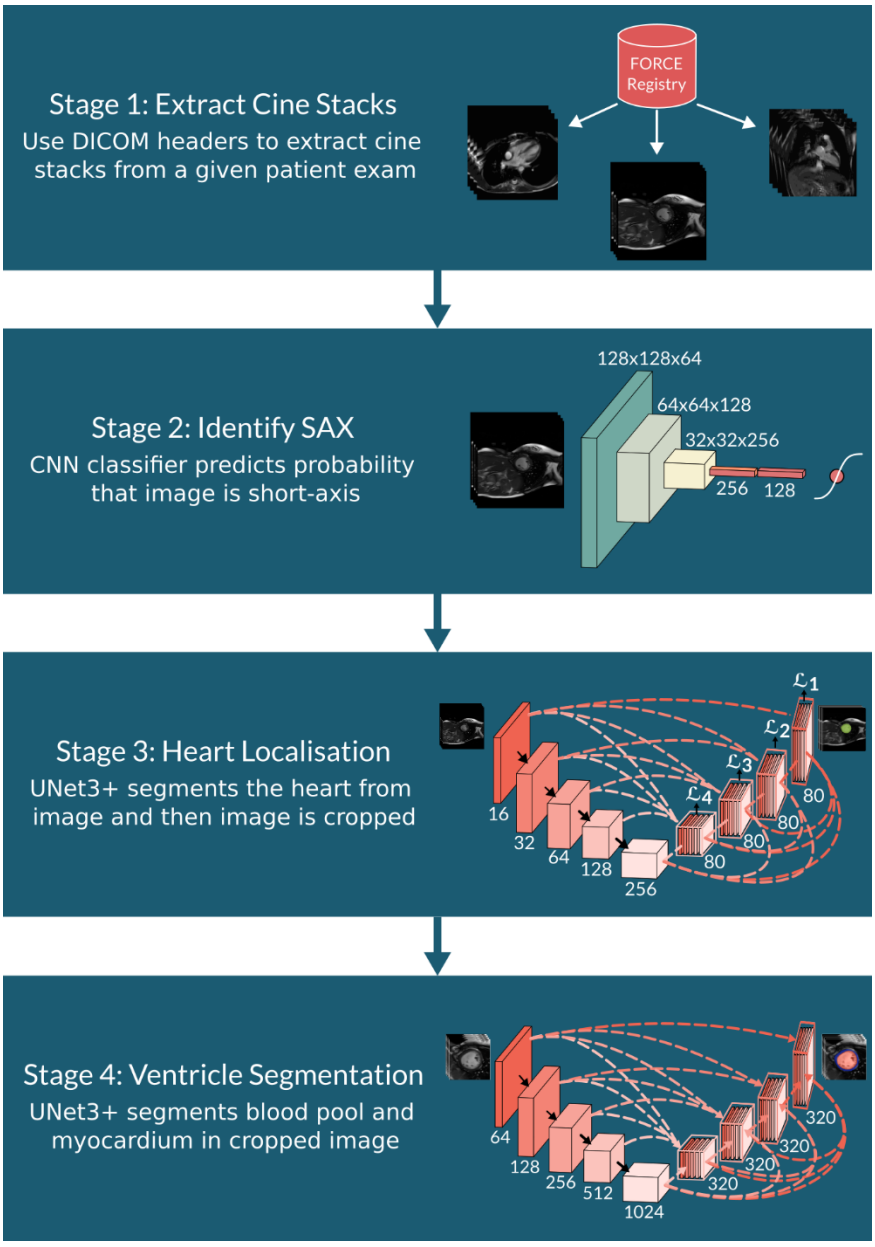



**Figure 1:** Outline of the 4 stages involved in the deep learning pipeline. DICOM = Digital Imaging and Communications in Medicine, FORCE = Fontan Outcomes Registry Using CMR Examinations, SAX = short-axis.

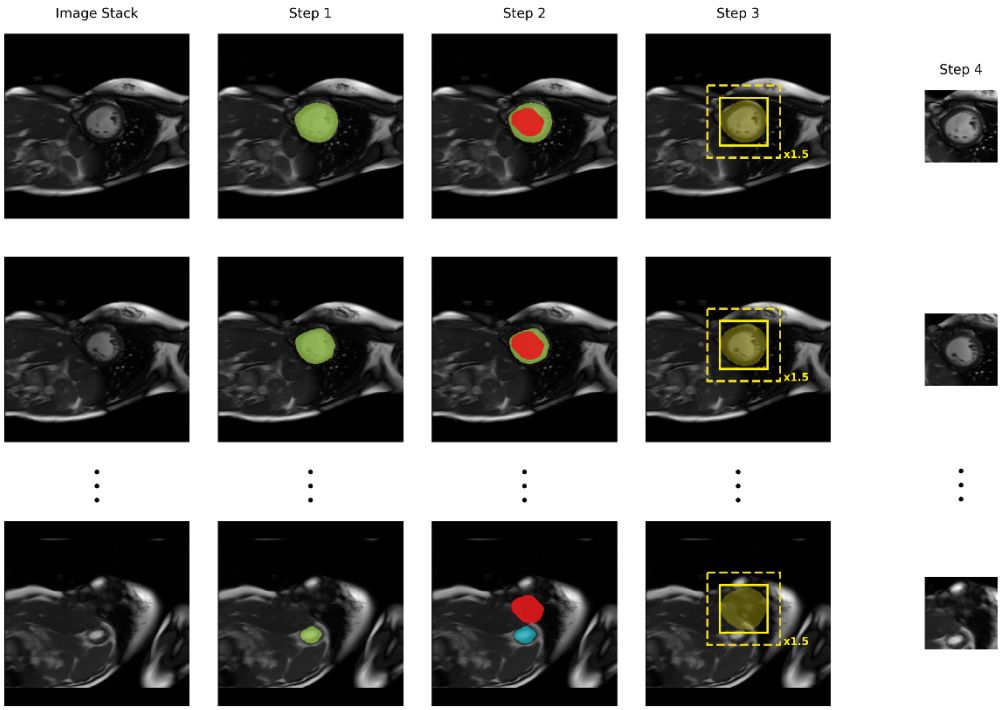



**Figure 2:** Schematic of the image cropping process. Step 1: Predicted segmentation mask based on epicardial border (green). Step 2: Removal of misclassified 'islands' (blue) that do not overlap with the intersection of all segmentations throughout the stack (red). Step 3: Creation of a box (solid yellow line) that bounds the union of all segmentations throughout the stack (yellow) and creation of the final bounding box (dashed yellow line). Step 4: Final cropped images using the bounding box from Step 3.

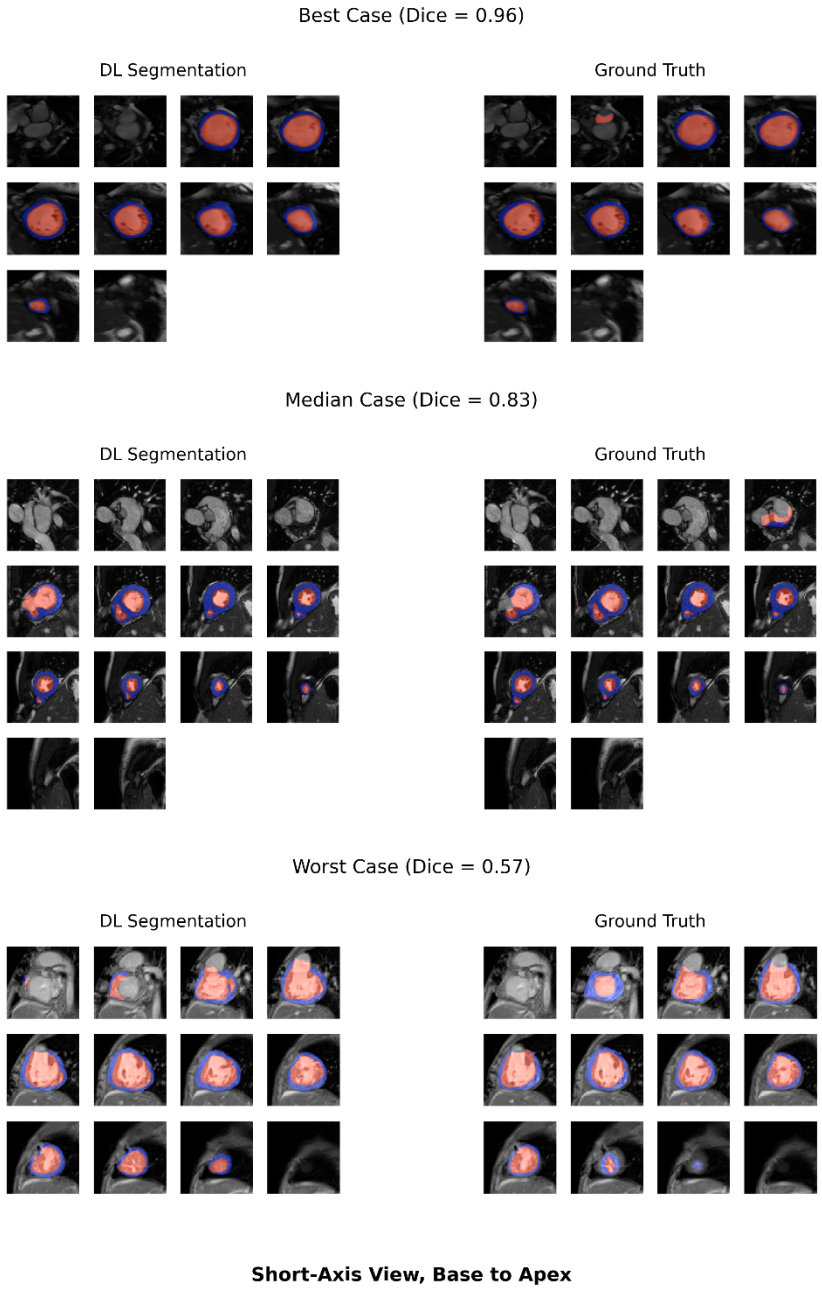



**Figure 3:** Deep learning (DL) segmentations versus ground truth segmentations for the best, median and worst test cases based on Dice score. Blood pool is shown in red and myocardium in blue.

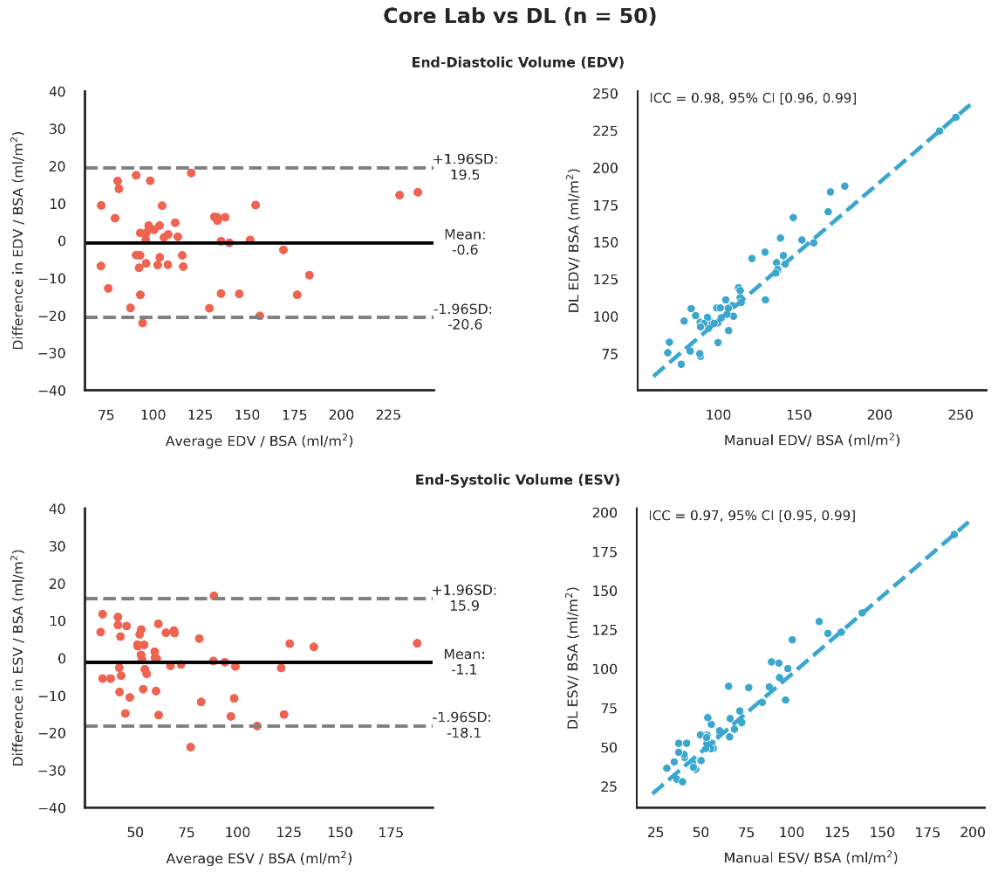



**Figure 4:** Bland-Altman plots and line of equality plots with intraclass correlation coefficients (ICC) comparing the manually-derived versus deep learning (DL)-derived end-diastolic volume (EDV) and end-systolic volume (ESV). BSA = body surface area.

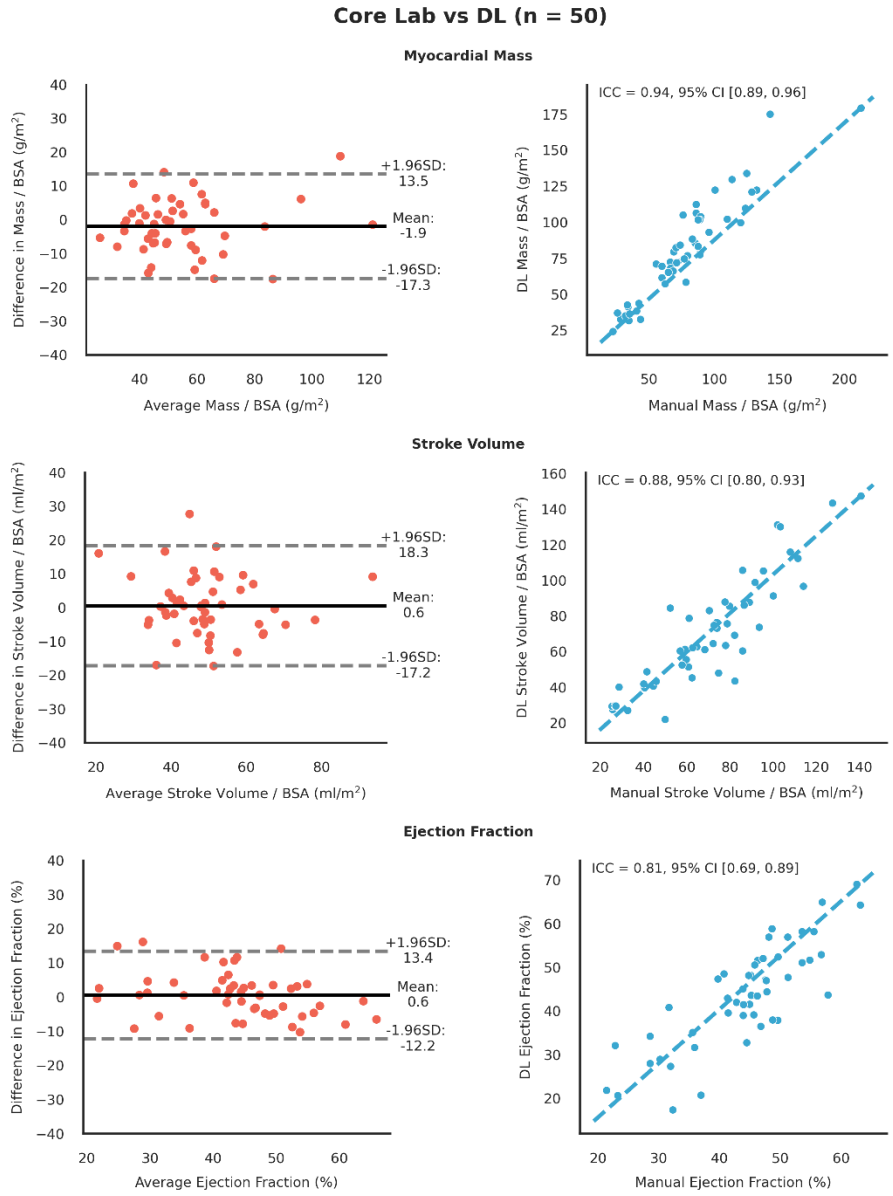



**Figure 5:** Bland-Altman plots and line of equality plots with intraclass correlation coefficients (ICC) comparing the manually-derived versus deep learning (DL)-derived ventricular mass, stroke volume and ejection fraction. BSA = body surface area.

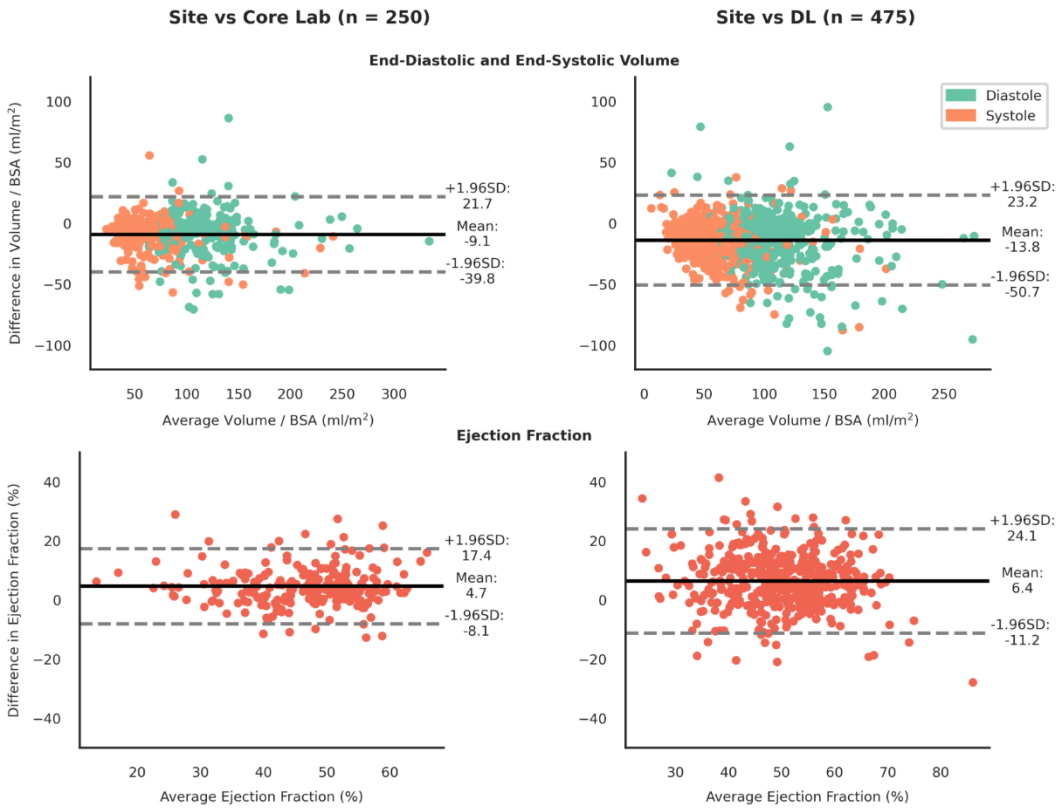



**Figure 6:** Bland-Altman plots comparing ventricular volume and ejection fraction. (left) Comparison between volumes ejections fraction from clinical reports entered by the contributing sites versus manual core laboratory calculated volumes and ejection fractions for 250 test patients. (right) Comparison between site-entered volumes and ejection fractions versus deep learning (DL)-derived volumes and ejection fractions for 475 test patients. BSA = body surface area.

## Table 1

### Demographics of the 250 Patients Used for Training and Validation of the Deep Learning Models and the 475 Patients Used for Testing the Deep Learning Pipeline

| | Number of Patients in the Manually Segmented Dataset (%) *n* = 250 | Number of Patients in the External Pipeline Test Set (%) *n* = 475 |
|---|---|---|
| Age | | |
| Adult | 137 (55%) | 212 (45%) |
| Child (< 16 years) | 113 (45%) | 263 (55%) |
| Median (IRQ), y | 16 (11–22) | 15 (9–19) |
| | | |
| Sex | | |

| | | | |
|---|---|---|---|
| | Male | 154 (61%) | 276 (58%) |
| | Female | 97 (39%) | 202 (42%) |
| Ventricle Systemic Circulation | | | |
| | Both | 133 (55%) | 238 (50%) |
| | RV Only | 64 (26%) | 137 (29%) |
| | LV Only | 47 (19%) | 100 (21%) |
| Scanner Field Strength | | | |
| | 1.5 T | 237 (95%) | 458 (96%) |
| | 3.0 T | 13 (5%) | 17 (4%) |

Note.—Unless otherwise indicated, data are presented as number of patients (percentage). LV = left ventricle, RV = right ventricle.

## Table 2

## Results for Body Surface Area-indexed End-diastolic Volume, End-systolic Volume, Ventricular Mass, Stroke Volume, and Ejection Fraction Obtained through the Deep Learning Pipeline and Manual Measurements for 50 Test Patient Examinations

| | Bias | Limits of Agreement | *P* Value | Intraclass Correlation Coefficient | Dice Score |
|---|---|---|---|---|---|
| End-Diastolic Volume (mL/m$^2$) | −0.6 | −20.6–19.5 | 0.75 | 0.98 (95% CI: 0.96, 0.99) | 0.91 (IQR, 0.89–0.94) |
| End-Systolic Volume (mL/m$^2$) | −1.1 | −18.1–15.9 | 0.56 | 0.97 (95% CI: 0.95, 0.99) | 0.86 (IQR, 0.82–0.89) |
| Ventricular Mass (g/m$^2$) | −1.9 | −17.3–13.5 | 0.12 | 0.94 (95% CI: 0.89, 0.96) | 0.74 (IQR, 0.70–0.77) |
| Stroke Volume (mL/m$^2$) | 0.6 | −17.2–18.3 | 0.85 | 0.88 (95% CI: 0.80, 0.93) | — |
| Ejection Fraction (%) | 0.6 | −12.2–13.4 | 0.52 | 0.81 (95% CI: 0.69, 0.89) | — |

Note.—Comparative evaluation using Bland-Altman, paired *t* test, intraclass correlation coefficient (95% CI), and Dice score (IQR).

## Table 3

## Median, IQR, and Range of the Time Taken and Number of Images Processed for Each Stage of the Deep Learning Pipeline for the Test Set (*n* = 475) Using One NVIDIA GeForce RTX 3090 GPU

| Stage | Variable | Median (IQR) | Range |
|---|---|---|---|
| 1: Extract Cine Stacks | | | |
| | Number of files | 3300 (1900–7100) | 460–21,000 |
| | Number of series | 35 (23–58) | 5–290 |
| | Number of stacks | 2 (1–3) | 1–16 |
| | Time Taken (s) | 9 (5–12) | 1–73 |
| 2: Identify SAX | | | |
| | Number of images classified | 10 (5–15) | 5–80 |
| | Time Taken (s) | 0.40 (0.20–0.48) | 0.030–1.45 |
| | Time per image (s) | 0.080 (0.040–0.096) | 0.0060–0.2900 |
| 3: Heart Localization | | | |
| | Number of images used for | 12 (12–14) | 7–25 |

| | | | |
|---|---|---|---|
| | localization | | |
| | Time Taken (s) | 1.5 (1.4–1.7) | 1.2–5.7 |
| | Time per image (s) | 0.12 (0.11–0.13) | 0.073–2.200 |
| 4: Ventricle Segmentation | | | |
| | Number of images segmented | 340 (240–360) | 100–1600 |
| | Time taken (s) | 11 (9–12) | 4–61 |
| | Time per image (s) | 0.032 (0.031–0.034) | 0.028–0.068 |
| Total time taken (s) | | 26 (21–32) | 13–110 |

Note.—Values are rounded to 2 significant figures.

## Table 4

### Labels Given to the End-diastolic and End-systolic Segmentation Outputs of the Deep Learning Pipeline for the Test Set (*n* = 475)

| | | Contains Artifact and/or Has Poor Image Quality | |
|---|---|---|---|
| Segmentation Quality Label | All | No | Yes |
| Satisfactory | 323 (68%) | 230 (74%) | 93 (57%) |
| Minor adjustments | 124 (26%) | 73 (23%) | 51 (31%) |
| Major adjustments | 26 (5%) | 8 (3%) | 18 (11%) |
| Crop fail | 2 (0.4%) | 0 (0%) | 2 (1%) |
| Total number | 475 | 311 | 164 |

Note.—Data are presented as number (percentage).

# Appendix E1. Distribution of Patients by Site Hospital

## Table E1

### Distribution of the 250 Patient Examinations by Site Hospital from the FORCE Registry Used in the Training, Validation and Testing of the Deep Learning Models in the Pipeline

| Hospital Name | Number of Patients (%), *n* = 250 | | | |
|---|---|---|---|---|
| | All | Training | Validation | Test |
| Boston Children's Hospital | 73 | 65 (89%) | 3 (4%) | 5 (7%) |
| Children's Healthcare of Atlanta | 10 | 4 (40%) | 2 (20%) | 4 (40%) |
| Children's Hospital of Philadelphia | 23 | 16 (70%) | 3 (13%) | 4 (17%) |
| Children's Hospital of Pittsburgh | 32 | 25 (78%) | 3 (9%) | 4 (12%) |
| Columbia University Irving Medical Center | 6 | 2 (33%) | 0 (0%) | 4 (67%) |
| Great Ormond Street Hospital | 1 | 1 (100%) | 0 (0%) | 0 (0%) |
| Lurie Children's Hospital | 16 | 10 (62%) | 2 (12%) | 4 (25%) |
| Mott's Children's Hospital | 14 | 8 (57%) | 2 (14%) | 4 (29%) |
| Nationwide Children's Hospital | 15 | 9 (60%) | 2 (13%) | 4 (27%) |
| Stollery Children's Hospital | 9 | 2 (22%) | 3 (33%) | 4 (44%) |
| Texas Children's Hospital | 33 | 25 (76%) | 3 (9%) | 5 (15%) |

| | | | | |
|---|---|---|---|---|
| Toronto Sick Kids | 12 | 6 (50%) | 2 (17%) | 4 (33%) |
| Yale New Haven Children's Hospital | 6 | 2 (33%) | 0 (0%) | 4 (67%) |

**Table E2**

**Distribution of the Patient Examinations by Site Hospital from the FORCE Registry**

| Hospital Name | Number of Patients (%) | |
|---|---|---|
| | *n* = 250 | *n* = 475 |
| Arkansas Children's Hospital | — | 1 (0%) |
| Boston Children's Hospital | 73 (29%) | 100 (40%) |
| Children's Healthcare of Atlanta | 10 (4%) | 54 (22%) |
| Children's Hospital of Philadelphia | 23 (9%) | 141 (56%) |
| Children's Hospital of Pittsburgh | 32 (13%) | 14 (6%) |
| Children's National Hospital | — | 1 (0%) |
| Columbia University Irving Medical Center | 6 (2%) | — |
| Great Ormond Street Hospital | 1 (< 1%) | 31 (12%) |
| Lurie Children's Hospital | 16 (6%) | 3 (1%) |
| Mott's Children's Hospital | 14 (6%) | — |
| Mount Sinai Medical Center | — | 10 (4%) |
| Nationwide Children's Hospital | 15 (6%) | 29 (12%) |
| Oklahoma University Health Sciences Center | — | 2 (1%) |
| Seattle Children's Hospital | — | 1 (0%) |
| Stollery Children's Hospital | 9 (4%) | — |
| Texas Children's Hospital | 33 (13%) | 36 (14%) |
| Toronto Sick Kids | 12 (5%) | 47 (19%) |
| Vanderbilt University Medical Center | — | 2 (1%) |
| Yale New Haven Children's Hospital | 6 (2%) | 3 (1%) |

*n* = 250 is the dataset used for the deep learning models. *n* = 475 is the external unseen test set.

# Appendix E2. SAX Identification Model

## Image Preprocessing

For the best classification performance, we ensured that the training data for the classifier was balanced between the non-SAX and SAX cines. The non-SAX cines were selected at random for a given examination providing a mix of different types of image views.

Images were resized to 128 × 128 using nearest-neighbor interpolation. Images were randomly augmented and then standardized such that, for each image, the mean was 0 and the standard deviation was 1.

The number of images used in total was 2500, split between training/validation/testing as 1750/250/500.

## Hyperparameter Optimization

The hyperband optimization was set to find the minimum binary cross-entropy loss. The following hyperparameters were optimized where the values in the brackets were the trial parameters, and the values in bold were the optimized parameter:

- Base Filters = [16,32,**64**]
- Kernel Size = [**3**,5]
- Dropout Rate = 0.0 - 0.5 (**0.2**)
- Second Dropout Layer = **True** or False
- Depth = [**1**,2,3]
- Layers = [2,**3**]

### Network Architecture

The SAX ID model was a CNN classifier that consisted of a stack of three 2D convolution and maxpooling layers, followed by two fully connected layers with a dropout layer following each fully connected layer and, finally, a sigmoid activation.

The three convolution layers used 64, 128 and 256 filters, respectively. The two fully connected layers had 256 and 128 nodes. The two dropout layers had a probability of 0.2. The final sigmoid activation layer outputted a probability of whether the input image was a SAX orientation image.

### Training Parameters

The classifier was trained with a binary cross-entropy loss. The model was trained with an Adam optimizer at a learning rate of 0.0001. The model was trained with a batch size of 128 until the loss stopped decreasing for 10 epochs.

## Appendix E3. Heart Localization Model

### Image Preprocessing

Images were resized such that their longest side was 256 pixels. Any nonsquare images were then zero-padded to a square such that all the input images were 256 × 256. Images were randomly augmented and then standardized.

The number of images used in total was 3303, split between training/validation/testing as 2312/331/660.

### Data Augmentation

Images were randomly augmented “on-the-fly” during the training using the Albumentations Python library.

- Random brightness between -10 - 10% of the image, with 30% probability

- Random contrast between -50 - 50% of the image, with 30% probability
- Random scaling between -30 - 30% of the image, with 100% probability
- Random translation between -10 and 10% of the image, with 100% probability
- Random rotation between -45 and 45 degrees, with 30% probability
- Random 90-degree rotation, with 20% probability
- One of:
  - Random image compression, with 80% as the lower bound for image quality and 100% as the upper bound, with 70% probability
  - Gaussian noise, with 10% probability
  - Motion blur, with 30% probability

## Hyperparameter Optimization

The objective of the hyperband optimization for this model was set to minimize the "blood pool pixel difference," which is a proxy for the volume. The blood pool pixel difference is the absolute difference between the number of pixels labeled as blood pool between the ground truth and the prediction.

- Loss = [**IoU**, Dice, Focal Tversky, Tversky, Categorical Crossentropy]
- Weighted Loss = **True**/False
- Deep supervision = **True**/False
- Dropout Rate = 0.0 - 0.5 (**0.5**)
- Depth = [**2**, 3]
- Kernel Size = [3, **5**]
- Base Filters = [**16**, 32, 64]

## Network Architecture

The heart localization model used the UNet 3+ architecture with 16 base filters, a weighted loss, deep supervision, a dropout rate of 0.5 and a final softmax layer.

Deep supervision is when segmentation masks are created at each decoding level, and Loss is calculated at each level. Weighted loss is how much each decoding level contributes to the total Loss. The weighted loss for the UNet 3+ was [0.25, 0.25, 0.25, 0.25, 1] for the decoding levels.

## Training Parameters

The classifier was trained with an IoU loss. The model was trained with an Adam optimizer at a learning rate of 0.0001. The model was trained with a batch size of 8 until the loss stopped decreasing for 10 epochs.

# Appendix E4. Ventricle Segmentation Model

## Image Preprocessing

The model takes square-cropped images, which are all resized to 128 × 128 pixels. Images were randomly augmented and then standardized.

The number of images used in total was 5512, split between training/validation/testing as 4308/362/842.

## Data Augmentation

For data augmentation, the image stack is first treated as a 3D volume. The Volumentations Python library was used to provide a random 90-degree rotation of the volume with a probability of 20%. Random rotation between-45 and 45 degrees with a probability of 40% was calculated using Scipy.

After initial "volume-wise" augmentation, a square bounding box containing the union of all masks in the slice direction was calculated for the augmented image stack. The bounding box is then expanded by a random value between 40%–60%. This mimics the output of the heart localization model, which finds a bounding box around the union of the predicted segmentation masks and expands the bounding box by 50%. The random scale factor also provides data augmentation.

After cropping, the images in the image stack were treated as separate 2D images. These 2D images were randomly augmented "on-the-fly" during the training using the Albumentations Python library.

- Random brightness between -10 - 10% of the image, with 30% probability
- Random contrast between -50 - 50% of the image, with 30% probability
- Random translation between -10 and 10% of the image, with 100% probability
- One of:
  - Random image compression, with 80% as the lower bound for image quality and 100% as the upper bound, with 70% probability
  - Gaussian noise, with 10% probability
  - Motion blur, with 30% probability

## Hyperparameter Optimization

Similar to the heart localization model, the hyperband optimization was set to minimize the "blood pool pixel difference."

## Network Architecture

The ventricle segmentation model used the UNet 3+ architecture with 64 base filters. This model did not use a weighted loss or deep supervision. It used a dropout rate of 0.3 and a final softmax layer.

- Loss = [IoU, Dice, Focal Tversky, **Tversky**, Categorical Crossentropy]
- Weighted Loss = True/**False**
- Deep Supervision = True/**False**
- Dropout Rate = 0.0 - 0.5 (**0.3**)
- Depth = [2, **3**]
- Kernel Size = [3, **5**]
- Base Filters = [16, 32, **64**]

## Training Parameters

The classifier was trained with a Tversky loss. The model was trained with an Adam optimizer at a learning rate of 0.0001. The model was trained with a batch size of 16 until the loss stopped decreasing for 10 epochs.

## Appendix E5. Qualitative Scoring

The pipeline was run on a further 475 patient examinations. The predicted end-diastolic and end-systolic segmentations for these patient examinations were qualitatively scored according to four categories:

1. Satisfactory segmentation - the predicted segmentation masks need no manual adjustments (Fig E1)
2. Minor adjustments - the masks need a small number of manual adjustments in 1-2 slices (Fig E1)
3. Major adjustments the masks need a large number of manual adjustments in >2 slices (Fig E1)
4. Crop failure - the heart localization stage failed to crop the location of the heart (Fig E2)

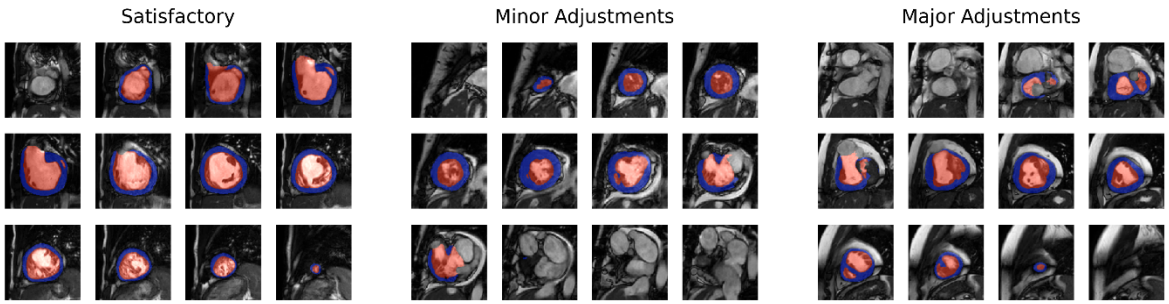



**Figure E1:** Examples of segmentations for the first three qualitative scoring criteria. The images were also labeled as having either good or poor image quality. Poor image quality is described as having poor resolution, poor contrast, containing an artifact, or having significant motion blur or noise, examples shown in Figure E2.

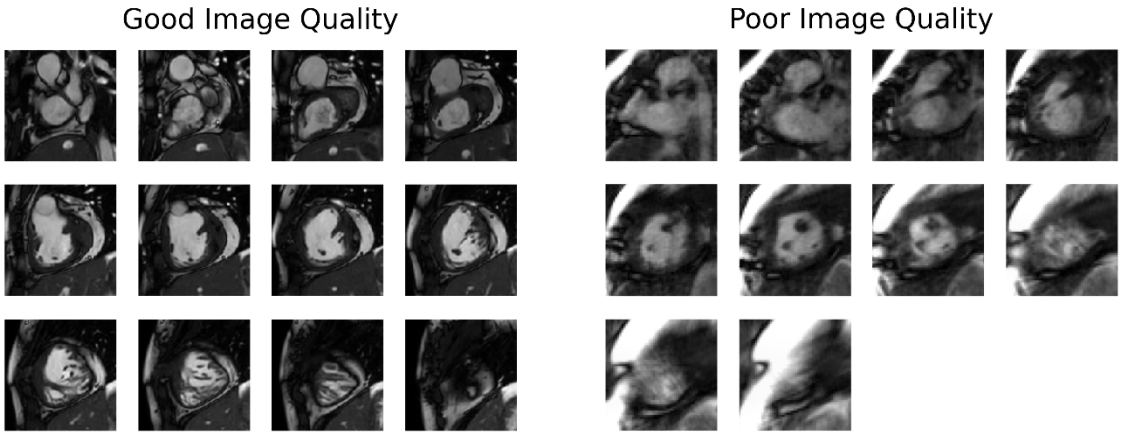



**Figure E2:** Examples of scans classed as having good and poor image quality. The images are cropped around the heart. All of the end-diastolic and end-systolic predictions (950 in total) were shown to the clinical observer individually and at random during the scoring. If the end-diastolic and end-systolic predictions were given different scores, the worst of the two scores was taken as the result of the given patient examination. A given patient examination was rated as having poor image quality if either the end-diastolic or end-systolic image was labeled as having a poor image quality.

## Appendix E6. Demographic Characteristics Comparison between Training, Validation and Test datasets

### Table E3

### Demographic Characteristics in the Training, Validation, and Test Datasets for Deep Learning Models are Compared

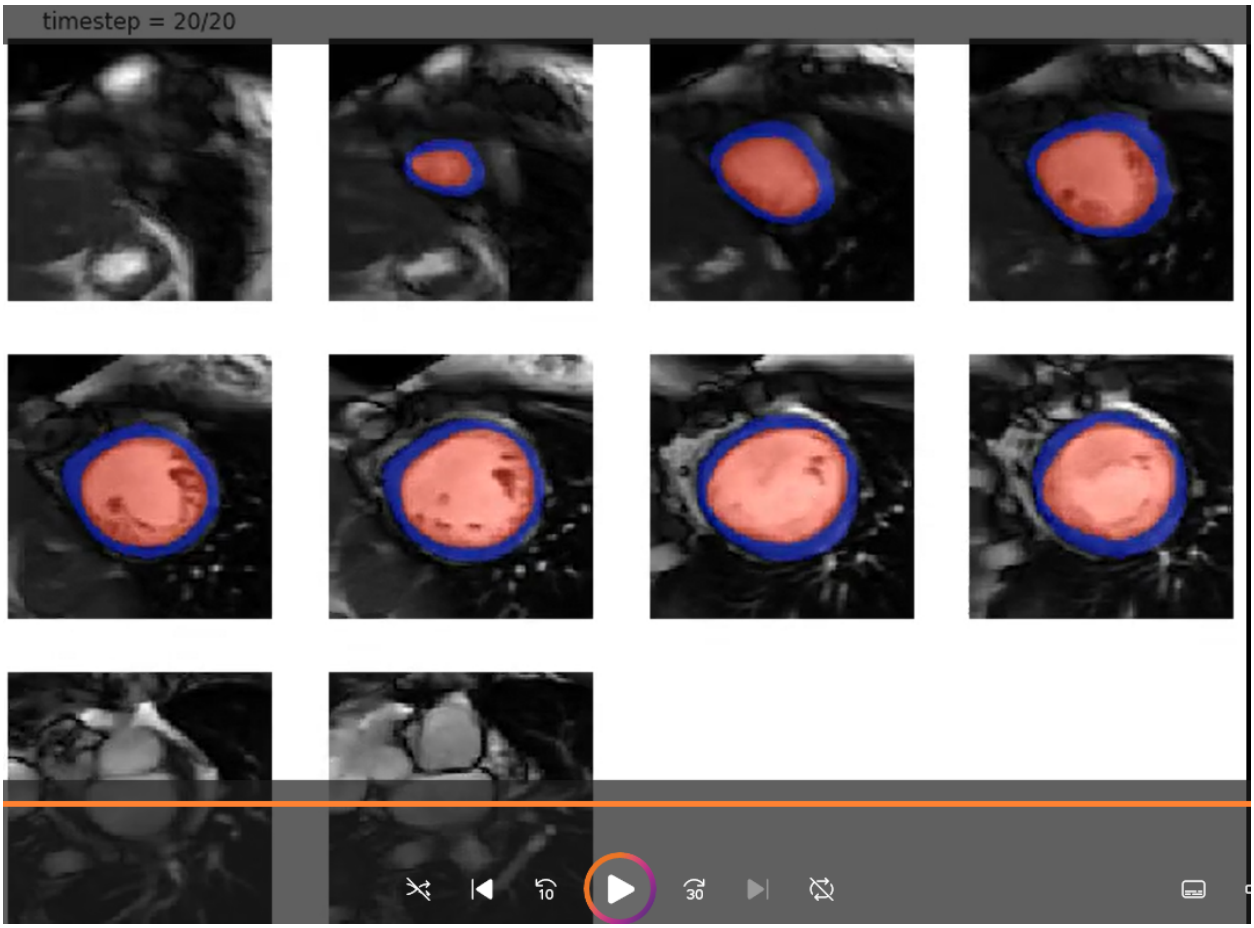
| | | Data Set | | |
|---|---|---|---|---|
| | | Training | Validation | Test |
| Age | | | | |
| | Median | 16.5 | 14.0 | 17.0 |
| | IQR | 12–22 | 11–19 | 10–21 |
| | Kruskal-Wallis Statistic | 0.95 | | |
| | *P* value | 0.62 | | |
| Body Surface Area ($m^{2)}$ | | | | |
| | Median | 1.59 | 1.52 | 1.66 |
| | IQR | 1.26–1.83 | 1.24–1.78 | 1.37–1.93 |
| | Kruskal-Wallis Statistic | 0.45 | | |
| | *P* value | 0.80 | | |
| Ventricle Systemic Circulation | | | | |
| | $X^2$ test Statistic | 1.7 | | |

| | | |
|---|---|---|
| | *P* value | 0.79 |
| | DOF | 4 |

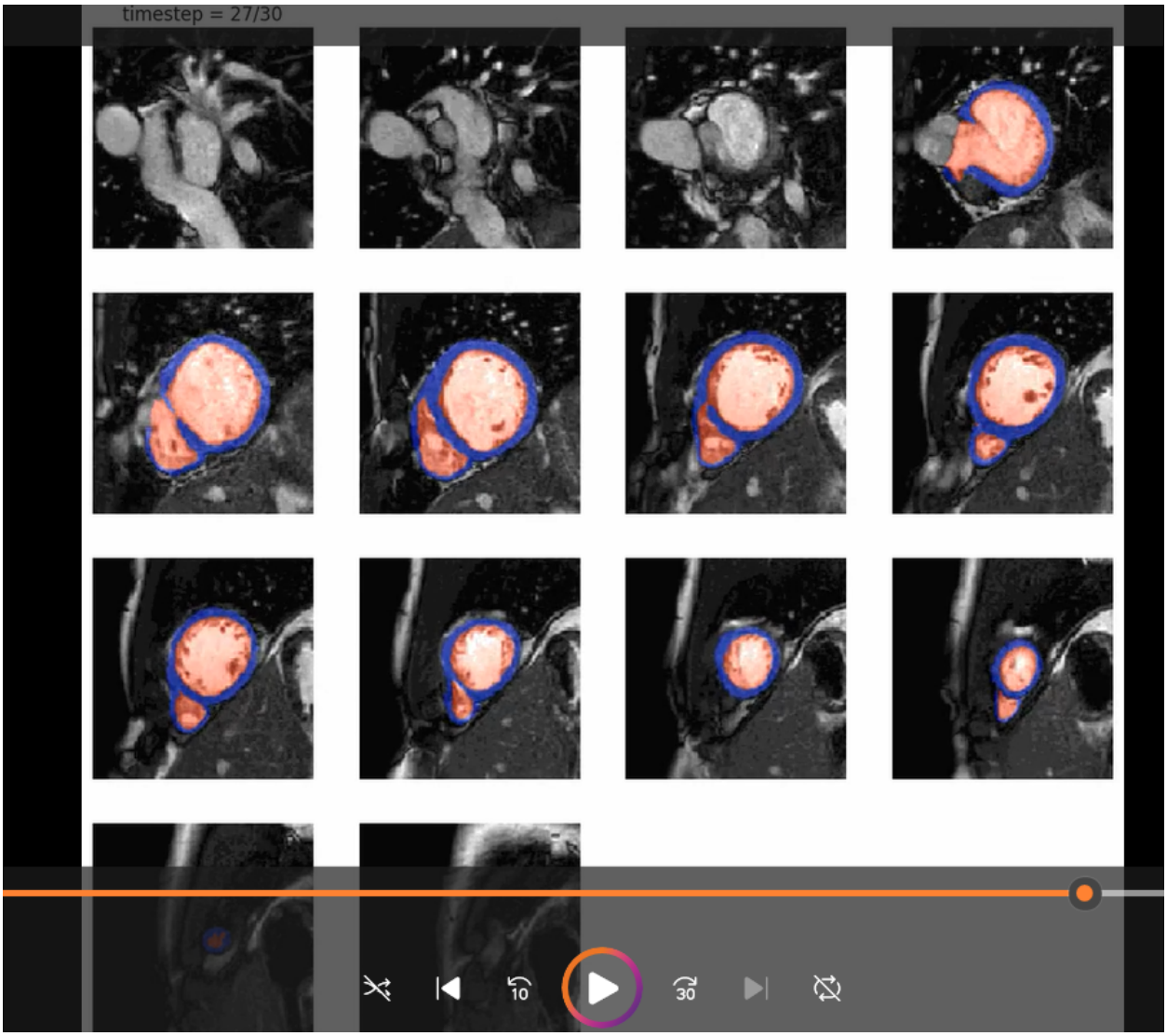
Note.—Age, BSA, and ventricular morphology types are tested for statistical significance. *P* values < .05 indicate significant differences.

## Appendix E7. Pipeline Segmentation Results over Time

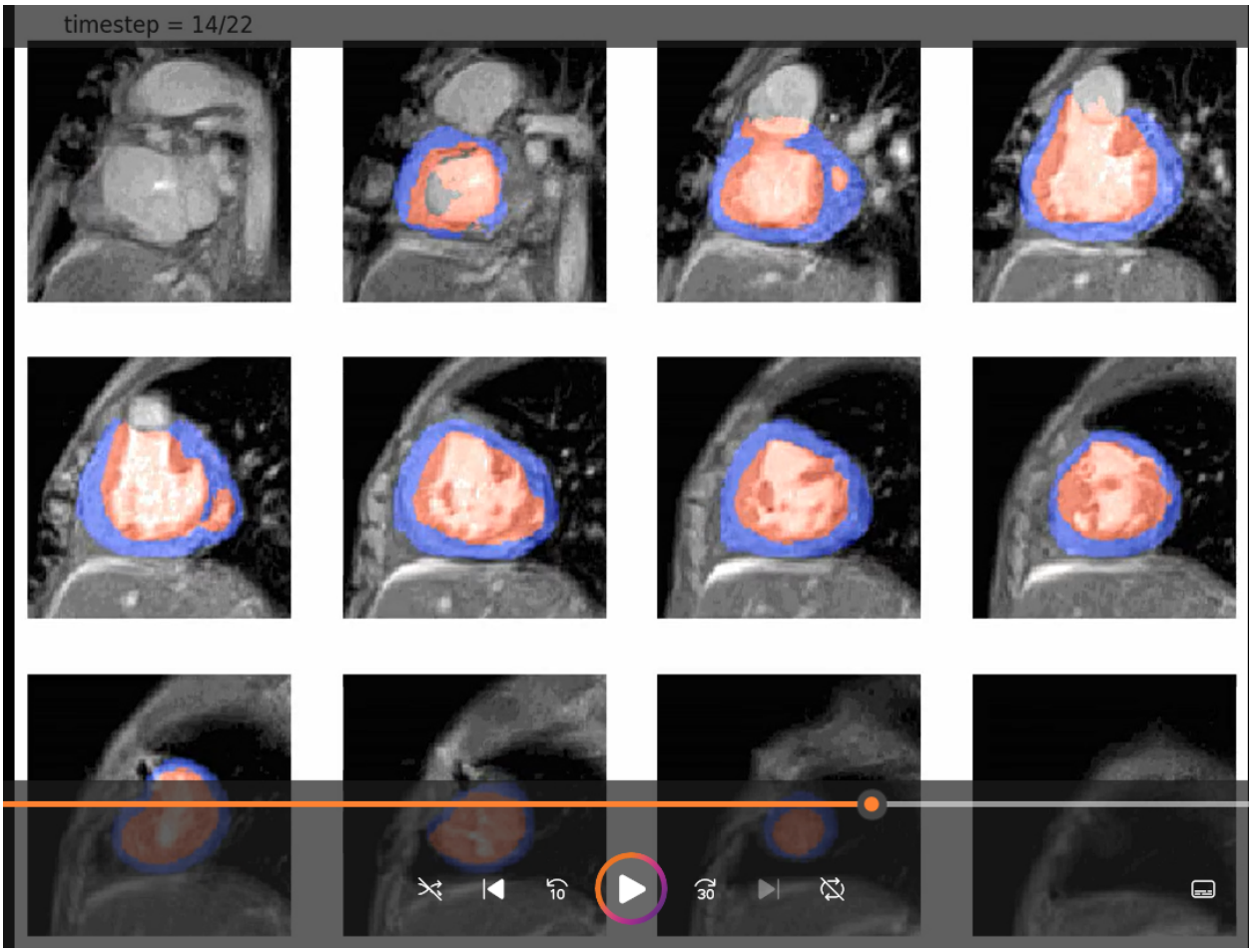
Movies 1–3 (multimedia) show the pipeline's segmentation outputs over time for the best, median and worst cases according to Dice score, corresponding to Figure 3 in the main paper.




**Movie 1:** Segmentation over time for the best case.




**Movie 2:** Segmentation over time for the median case.




**Movie 3:** Segmentation over time for the worst case.

# A Deep Learning Pipeline for Assessing Ventricular Volumes from a Cardiac MRI Registry of Patients with Single Ventricle Physiology

**Key Result**

The developed deep learning (DL) pipeline can provide standardized segmentation for patients with single ventricle physiology that is robust to highly variable anatomy and heterogeneous data collected from multiple centers.

## Patients:

Training, Validation, & Internal Test Sets: 250 cardiac MRI examinations from 13 centers (FORCE registry)

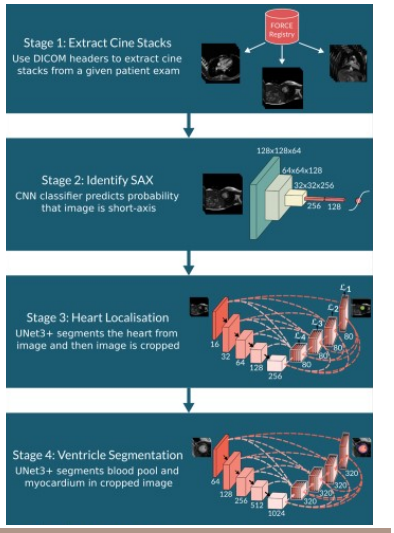
External Test Set: 475 unseen cardiac MRI examinations from FORCE registry

## Methods:

- The pipeline contained three DL models (see Figure).
- DL-derived volumetric and functional metrics were compared with ground truth manual segmentation-derived measurements.
- The pipeline was qualitatively evaluated on the external test set.

## Results:

- Median (IQR) Dice scores achieved by the pipeline:
  - End-diastolic volume = 0.91 (0.89–0.94)
  - End-systolic volume = 0.86 (0.82–0.89)
  - Myocardium = 0.74 (0.70–0.77)
- Segmentation quality on external test set examinations: 68% were satisfactory, 26% needed minor adjustments, 5% needed major adjustments, and 0.4% had cropping model failure.




Yao T and St. Clair N et al. Published Online: November 15, 2023
https://doi.org/10.1148/ryai.230132

Radiology: Artificial Intelligence