

1 **Deep Learning vs manual techniques for assessing left ventricular ejection fraction in 2D**  
2 **echocardiography: validation against CMR**

3 **Short title:** Validation of DL LVEF echocardiography vs CMR.

4  
5 Eric Saloux<sup>a,b,\*¶</sup>, Alexandre Popoff<sup>c</sup>, Hélène Langet<sup>d</sup>, Paolo Piro<sup>c</sup>, Camille Ropert<sup>a</sup>, Romane  
6 Gauriau<sup>c</sup>, Romain Stettler<sup>a</sup>, Mihaela Silvia Amzulescu<sup>e</sup>, Guillaume Pizaine<sup>c</sup>, Pascal Allain<sup>c</sup>,  
7 Olivier Bernard<sup>f</sup> Amir Hodzic<sup>a</sup>, Alain Manrique<sup>a,b</sup>, Mathieu De Craene<sup>c¶</sup>, Bernhard L. Gerber<sup>c</sup>

8  
9 a) Centre Hospitalier Universitaire de Caen Normandie, France

10 b) EA 4650, Caen University, FHU REMOD-VHF, France

11 c) Philips Research, Medical Imaging (Medisys), Suresnes, France

12 d) Philips Clinical Research Board, Suresnes, France

13 e) Cliniques Universitaires Saint-Luc UCL, Brussels, Belgium

14 f) CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, University of Lyon 1,  
15 Villeurbanne, France

16  
17 **Subject Terms:** Transthoracic Echocardiography,

18 **Relation with Industry Disclosure:** AP, MDC, GP and PA are employed by Philips Medical  
19 Systems. HL, RG and PP were employed by Philips Medical Systems at the time of the study.  
20 Centre Hospitalier Universitaire Caen Normandie and Cliniques Universitaires Saint-Luc have  
21 a master research agreement with Philips Medical Systems.

22 **Corresponding author:**

23 Email: [saloux-e@chu-caen.fr](mailto:saloux-e@chu-caen.fr)

24

25 **NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

## 26 **Structured Abstract**

27 **Objective:** To evaluate accuracy and reproducibility of 2D echocardiography (2DE) left  
28 ventricular (LV) volumes and ejection fraction (LVEF) estimates by Deep Learning (DL) vs.  
29 manual contouring and against CMR.

30 **Background:** 2DE LV manual segmentation for LV volumes and LVEF calculation is time  
31 consuming and operator dependent.

32 **Methods:** A DL-based convolutional network (DL1) was trained on 2DE data from centre A,  
33 then evaluated on 171 subjects with a wide range of cardiac conditions (49 healthy) – 31  
34 subjects from centre A (18%) and 140 subjects from centre B (82%) – who underwent 2DE and  
35 CMR on the same day. Two senior (A<sub>1</sub> and B<sub>1</sub>) and one junior (A<sub>2</sub>) cardiologists manually  
36 contoured 2DE end-diastolic (ED) and end-systolic (ES) endocardial borders in the cycle and  
37 frames of their choice. Selected frames were automatically segmented by DL1 and two DL  
38 algorithms from the literature (DL2 and DL3), applied without adaptation to verify their  
39 generalizability to unseen data. Interobserver variability of DL was compared to manual  
40 contouring. All ESV, EDV and EF values were compared to CMR as reference.

41 **Results:** 50% of 2DE images were of good quality. Interobserver agreement was better by DL1  
42 and DL2 than by manual contouring for EF (Lin's concordance = 0.9 and 0.91 vs. 0.84), EDV  
43 (0.98 and 0.99 vs. 0.82), and ESV (0.99 and 0.99 vs. 0.89). LVEF bias was similar or reduced  
44 using DL1 (-0.1) vs. manual contouring (3.0), and worse for DL2 and DL3. Agreement between  
45 2DE and CMR LVEF was similar or higher for DL1 vs. manual contouring (Cohen's kappa =  
46 0.65 vs. 0.61) and degraded for DL2 and DL3 (0.48 and 0.29).

47 **Conclusion:** DL contouring yielded accurate EF measurements and generalized well to unseen  
48 data, while reducing interobserver variability. This suggests that DL contouring may improve  
49 accuracy and reproducibility of 2DE LVEF in routine practice.

50 **Keywords**

51 Transthoracic Echocardiography, magnetic resonance imaging, validation, deep learning

52

## 54 **List of Abbreviations and Acronyms**

AP4	apical 2-chamber
AP2	apical 4-chamber
CCC	Lin's concordance correlation coefficient
CMP	cardiomyopathy
DL	deep learning
EF	ejection fraction
ED[V]	end-diastolic [volume]
ES[V]	end-systolic [volume]
LV	left ventricle
VOL	healthy volunteer

55

## 56 **Introduction**

57 Accurate and reproducible echocardiographic assessment of left ventricular (LV) volumes and  
58 ejection fraction (EF) is crucial in clinical decision-making and risk stratification (1–6). Hence  
59 LVEF thresholds are used for decision making in heart failure (7) and coronary artery (8) and  
60 valvular heart (9) diseases. Simpson’s method from two-dimensional (2D) 2- and 4-chamber  
61 views is currently the preferred approach for evaluation of LV volumes and EF by  
62 echocardiography. Yet it is time consuming, subject to wide interobserver variability (10,11),  
63 and the reproducibility is highly affected by various factors such as operator experience and  
64 image quality. Accordingly, 2D echocardiography has shown to be less reproducible than  
65 cardiac magnetic resonance (CMR)(12), which is currently considered the reference standard  
66 for evaluation of LVEF and volumes. This approach however suffers from higher costs and less  
67 frequent availability.

68 Deep learning (DL) allows automated contour detection offering the promise of faster and  
69 potentially more accurate and reproducible evaluation of LV volumes and EF by  
70 echocardiography (3,7,13,14). In this work, we developed a new DL algorithm for manual  
71 contouring based on a U-Net convolutional network architecture, using an anonymised database  
72 of echocardiographic images. The aim of the present study was to evaluate the generalizability  
73 and accuracy of this new algorithm relative to manual contouring and against cardiac MR  
74 volumes and EF as a reference. We evaluated our algorithm using a set of multimodal data from  
75 171 subjects from two centres. We also compared the automated contouring and resulting LV  
76 volumes to their manual counterparts, as obtained by different junior and senior observers  
77 across the two centres and used CMR as an independent 3D modality to evaluate differences in  
78 bias between manual and automated contouring. Finally, we benchmarked our DL algorithm  
79 against two other DL implementations from the literature and made the 2DE database, CMR

80 LVEF values and the Simpson bi-plane code publicly available for reproducibility of this paper  
81 and further benchmarks.

## 82 **Methods**

### 83 **Study Design**

84 The present study is a retrospective analysis of echocardiography and CMR data from  
85 participants previously enrolled in prospective trials performed either at Centre Hospitalier  
86 Universitaire de Caen Normandie (denoted as clinical centre A) or at Cliniques Universitaires  
87 Saint-Luc (denoted as clinical centre B), following approval by an ethic committee. The studies  
88 were approved by the IRB in charge (either Comité de Protection des Personnes Nord-Ouest  
89 III, Caen, France or Comité Ethique Hospitalo-Facultaire de l'Université Catholique de  
90 Louvain, Brussels, Belgium). The retrospective use of the data satisfies the European Union  
91 (EU) General Data Protection Regulation (GDPR) requirements. Data from participants who  
92 had undergone 2D echocardiography and cardiac CMR within 24 hours and who were found in  
93 sinus rhythm were analysed in the present study, resulting in a database of 171 subjects in total.  
94 Among those, there were 49 healthy volunteers and 122 patients with various cardiac  
95 pathologies: 37 with an ischaemic heart disease and a previous myocardial infarct, 35 with a  
96 non-ischaemic dilated cardiomyopathy, 39 with a valvular heart disease (including 19 aortic  
97 stenoses, 17 mitral regurgitations, 2 mitral repairs and 1 aortic regurgitation), and 11 with a  
98 hypertrophic cardiomyopathy. There was no prior selection based on image quality, thus  
99 reflecting a realistic range of echogenicity and artefacts in the database. For each subject,  
100 demographic and anthropometric data (age, sex, height, weight), diastolic and systolic blood  
101 pressures, and cardiovascular risk factors were also extracted.

### 102 **2D transthoracic echocardiography**

103 Standardized comprehensive transthoracic echocardiographic examinations had been acquired  
104 according to established guidelines (15) using Philips IE 33 and EPIQ 7 ultrasound system

105 equipped with a X5-1 transducer in harmonic imaging (Philips Medical Systems, Andover, MA,  
106 United States), and stored on a PACS server (Intellispace Cardiovascular (ISCV), Philips  
107 Medical Systems, Andover, MA, United States).

## 108 **Echocardiography measurements**

109 **Manual measurements** Echocardiographic images were anonymized, exported in DICOM  
110 format and analysed off-line. Three successive cardiac cycles were available for apical 4- and  
111 2-chamber views. Two senior ( $A_1$ : ES) and ( $B_1$ : BLG) and one junior ( $A_2$ : RS) cardiologists  
112 manually contoured the ED and ES endocardial borders in the cardiac cycle and on the image  
113 frames of their choice, while blinded to quantitative outcomes. Senior cardiologists had more  
114 than 20 years of experience in echocardiography, the junior cardiologist had 3 months of  
115 training. For DL analysis, a deep convolutional neural network (U-Net) model was used to  
116 segment the LV cavity on the frames selected by the three observers. The same biplane  
117 Simpson's method was used to compute all LV volumes and EF (see *Volumes computation*  
118 below). The study protocol is summarized in (Fig 1). Data were analysed in two independent  
119 ways (manual EF, and DL-EF for the 3 compared DL algorithms) for all frames selected by the  
120 3 observers. Observers were blinded to other manual and automated quantifications. Manual  
121 analysis was performed with a custom Python script running a Web browser-based interface to  
122 present images in random order and to save the contours. The observers first selected a cycle  
123 among the three consecutive cycles, then determined ED and ES within this cycle. An  
124 interactive graphical interface allowed the contouring of the cavity on both the AP2 and AP4  
125 views. All observers were instructed to respect the following conventions: i) include trabeculae  
126 and papillary muscles in the LV cavity; ii) keep consistency in excluding/including tissue  
127 between ED and ES; iii) terminate the contours in the mitral valve plane on the ventricular side  
128 of the bright ridge, at the points where the valve leaflets are hinging. Automated segmentation  
129 was performed using three deep learning algorithms (see below). The algorithms were run on

130 all frames selected by observers. Therefore, DL measurements will be presented systematically  
131 for the 3 observers (junior  $A_1$  and senior  $A_2$  and  $B_1$ ).

132 **Fig 1. Flowchart of the study.** The deep convolutional neural network (U-Net) model was  
133 trained on about 700 apical 2- and 4-chamber views from site A. The performance of Deep  
134 learning (DL) analysis using the U-Net model was then evaluated on data from 171 subjects  
135 from sites A and B, who underwent both transthoracic echocardiography and cardiac magnetic  
136 resonance (CMR). CMP = cardiomyopathies. DL = Deep Learning. ED = end-diastole. ES =  
137 end-systole.

138 **Volumes computation** A standard Simpson's rule for biplane EF computation was  
139 implemented according to (16) and applied to manual and DL segmentations. The code for this  
140 Simpson's implementation is made publicly available on a git repository. The first step is first  
141 to find for both the AP2 and AP4 images a rotated bounding box fully encompassing the input  
142 contour. The second step is to find the apex and mitral points. The apex is defined as the contour  
143 point the closest to the middle of the bounding box top edge. The mitral point is found as the  
144 contour point that intersects the long axis direction of the bounding box. Mitral and apex points  
145 define the long axis segment in both AP2 and AP4 images. These long axis segments are divided  
146 into twenty points at which two radiuses are cast towards both sides of the input contour. By  
147 summing the N ellipsoidal cylinders from the AP2 and AP4 contours, one obtains the EDV and  
148 ESV values.

149 **Image quality assessment** To assess the feasibility of DL with respect to image quality, senior  
150 observer  $B_1$  ranked the image quality of all frames (i.e., both ES and ED frames for AP2 and  
151 AP4) into three categories: good, meaning the endocardial wall was visible for the whole LV;  
152 fair, meaning the endocardium had to be visually interpolated at some locations; and poor,  
153 meaning that the frame was not of sufficient quality for being quantified. Senior observer  $A_1$   
154 also classified all DL segmentations according to whether he would have edited or not the



155 automated contouring. Finally,  $A_1$  also counted the cases for which DL outperformed Junior  
156 Observer  $A_2$ , assessing if DL was worse, as good or better than  $A_2$ 's contouring for  
157 quantification purposes.

### 158 **Deep learning algorithms**

159 A U-Net architecture was trained to segment the LV cavity mask using ED and ES manual  
160 contouring from senior cardiologist  $A_1$  on an independent database of about 700 anonymised  
161 apical AP2 and AP4 views from 237 subjects. That database consisted of patients with ischemic  
162 cardiomyopathy (59%), dilated cardiomyopathy (6%), valvular pathologies (9%), hypertrophic  
163 cardiomyopathy (3%). Remaining subjects in the training database underwent 2D echo because  
164 of arterial hypertension or other cardiovascular risk factors. For all included subjects, image  
165 quality had been deemed satisfactory by  $A_1$  to be manually contoured in the AP2 and AP4  
166 views. The convolutional neural network architecture followed the standard U-Net pattern,  
167 stacking elementary blocks of convolutional, activation and batch normalization layers (17).  
168 The network took as input the full scan-converted B-Mode image, resized to a 192x256 size.  
169 The activation of the final layer (sigmoid) outputted a continuous mask image that took values  
170 between 0 and 1. After thresholding this result at 0.5, the largest contour was extracted for the  
171 AP2 and AP4 ES and ED frames. This method is referred to as DL1 in the remainder of this  
172 paper.

173 For benchmarking DL1 against other techniques, two other DL implementations from the  
174 literature were evaluated. First, the e-Net architecture from Leclerc et al. (13) was applied on  
175 the same frames as the DL method. This method is referred to as DL2 in the remainder of this  
176 paper. The DL2 network was trained on GE images, on the CAMUS public database (13). The  
177 weights of the network were used as such, without any adaptation. Finally, the method of Zhang  
178 et al. (18) was applied similarly without any retraining. This model is referred to as DL3.

## 179 **Magnetic resonance imaging**

180 All subjects had undergone a standardized CMR myocardial function study on a 3T scanner  
181 (Achieva, Philips Medical Systems, Best, the Netherlands). The CMR exam was performed  
182 within a 24 hours window from the echocardiographic exam. 10-12 consecutive short axis  
183 images covering the entire LV, and respectively one 2- 3- and 4 chambers long-axis cine SSFP  
184 images were acquired for assessment of myocardial function. CMR RV and LV volumes and  
185 EF were computed using Segment version 2.2 (<http://segment.heiberg.se>) (19) or Medis  
186 software (Medical Imaging Systems, Leiden, the Netherlands) from short-axis cine images by  
187 semi-automatically contouring the endo- and epicardial contours in the end-diastolic (ED) and  
188 end-systolic (ES) phases. These quantifications were performed by an independent EuroCMR  
189 level III certified operator (MSA) blinded to the quantitative findings of echocardiographic  
190 operators. Papillary muscles and trabeculations were included as blood volume in the cavity  
191 contour.

## 192 **Statistical analysis**

193 Statistical analyses were performed using the epiR and psych R packages. Continuous variables  
194 are presented as mean values±SD, categorical variables as counts and percentages. Continuous  
195 variables were compared using the independent sample Student t test if normally distributed, or  
196 else using either the Wilcoxon signed rank test (paired data) or the Mann-Whitney (unpaired  
197 data) tests. A p-value  $p < 0.05$  was considered statistically significant. Agreements between 2D-  
198 echo DL and manual segmentations for each observer was assessed using the DICE similarity  
199 coefficient computed as  $S_{DICE} = 2 \cdot \frac{n\{E_{DL} \cap E_{manual}\}}{n\{E_{DL}\} + n\{E_{manual}\}}$  where  $E_{DL}$  and  $E_{manual}$  are the sets of pixels  
200 found within the LV cavity by DL and manual contouring, and  $n\{ \cdot \}$  is the number of pixels in  
201 a given set. DICE is a standard measure to compare the overlap of two binary segmentation  
202 masks. Both interobserver and echo vs. CMR agreements were measured with Lin's  
203 concordance correlation coefficient (CCC) for EDV, ESV, and EF. To better evaluate the

204 impact of different EF values using either DL vs. manual contouring or CMR vs. 2D echo, we  
 205 categorized EF into three thresholds:  $EF < 40\%$ ,  $40\% \leq EF \leq 50\%$ ,  $EF > 50\%$  matching the  
 206 guidelines for the LVEF stratification of HF patients (20). We then measured agreement to  
 207 classify subjects into these 3 groups among each observer and DL by 2D echo and CMR using  
 208 Cohen's kappa ( $\kappa$ ) coefficient.

## 209 Results

### 210 Study Population

211 Table 1 presents baseline characteristics of the study population. The validation cohort of this  
 212 study was composed of  $n=31$  (18%) patients from centre A and  $n=140$  (82%) patients from  
 213 centre B. The population had a wide range of LV ejection fraction and volumes. There were no  
 214 significant differences in hemodynamics between echo and CMR studies. As expected, there  
 215 were significant differences in age, EDV, ESV and EF among subjects with different cardiac  
 216 conditions. Image quality of echo images was rated by  $B_1$  as good (resp. fair and poor) for 50%  
 217 (resp. 42% and 8%) of the frames composing the dataset.

218 **Table 1. Patient population characteristics.**

	<b>HCM</b> N = 11	<b>ISCH</b> N = 37	<b>NON-ISCH</b> N = 35	<b>VALV</b> N = 39	<b>VOL</b> N = 49	<b>ALL</b> N= 171
<b>Age, y</b>	42±13	56±14 *	57±18 *	66±13 *	48±14	55±14
<b>Males, # (%)</b>	9 (82)	36 (97)	25 (71)	30 (77)	35 (71)	135 (79)
<b>BSA, m<sup>2</sup></b>	1.9±0.1	1.9±0.2	1.9±0.2	1.8±0.2	1.8±0.2	1.9±0.2
<b>HR bpm</b>	73±12	75±16 *	70±13 *	67±7	64±10	69±10
<b>DBP, mmHg</b>	71±9	69±13 *	76±9	73±10	79±11	74±12
<b>SBP, mmHg</b>	119±15	117±23	119±14	129±19	128±22	123±21
<b>CMR-EDV, mL</b>	177±53	233±93 *	264±69 *	199±66 *	150±33	204±79
<b>CMR-ESV, mL</b>	63±24	163±97 *	188±70 *	77±45	57±15	112±81
<b>CMR-EF, %</b>	65±9	35±15 *	31±13 *	63±12	62±6	50±19

219 Baseline, echo and CMR characteristics. HCM=hypertrophic cardiomyopathy, NON-  
220 ISCH=dilated non-ischaemic cardiomyopathy, ISCH=ischemic heart disease,  
221 VALV=valvular diseases, VOL=healthy volunteer. Body Surface Area (BSA) was calculated  
222 using the Mosteller formula. \* indicates p-values  $p < 0.05$  vs. the VOL subgroup.

### 223 **Feasibility of DL-based contouring**

224 DL computation using the DL1 network of EDV and ESV and EF was feasible in all subjects  
225 and images and took about 60 ms per image (including biplane Simpson's computation).  
226 Typical examples of DL1 contouring are shown in (Fig 2). Reviewer A<sub>1</sub> considered DL1  
227 segmentations as acceptable, and not requiring further editing, in 70% of the frames (64% for  
228 fair/poor images). A<sub>1</sub> also considered 16.4% of the frames to be better segmented by DL1 than  
229 manual segmentation by the junior observer A<sub>2</sub>. For the latter, A<sub>1</sub> was not blinded to which  
230 method was used to produce the segmentation result. In only 6 cases (3%) DL segmentation  
231 was considered to have failed as illustrated in (S1 Fig). In a patient with non-ischemic dilated  
232 CMP and a patient with mitral regurgitation, a hyper-intense valve apparatus in the image  
233 disrupted the DL1 segmentation and yielded a cavity segmentation with a highly irregular  
234 shape. In two patients (with aortic stenosis and mitral regurgitation), a partly non-visible  
235 endocardial wall impeded automatic segmentation. In the two last outliers, one hypertrophic  
236 and aortic stenosis, with poor LV function and low EF, local DL1 segmentation errors had a  
237 higher impact than in subjects with good LV function.

238 **Fig 2. Overlay of typical DL1 segmentations (yellow mask) and manual contouring by**  
239 **senior observer B<sub>1</sub> (red dotted contour) for four representative subjects.** From top to  
240 bottom: healthy volunteer (VOL), subject with a ischemic heart disease (ISCH), subject with  
241 a hypertrophic cardiomyopathy (HCM), and subject with an aortic stenosis (VALV). From left  
242 to right: end-diastolic frame in apical 4-chamber view, end-systolic frame in apical 4-chamber  
243 view, end-diastolic frame in apical 2-chamber view, and end-systolic frame in apical 2-chamber  
244 view.

245 **Agreement of DL and manual contouring.**

246 (Fig 3) shows the DICE values spread when comparing DL1 vs. manual contours on the (ES, ED  
247 ) × (AP2, AP4) frames of every observer. For all echo views and observers, the average DICE  
248 values were over 90%. However, a more elevated spread appeared in the junior (A<sub>2</sub>) compared  
249 to the two senior (A<sub>1</sub>, B<sub>1</sub>) observers. In addition, the AP4 view showed a higher consistency  
250 between manual and DL results. Similarly, within each echo view, lower DICE values were  
251 found in ES than in ED (with the exception of observer B<sub>1</sub> in the A2C view).

252 **Fig 3. DL1 vs. manual comparison.** Left-ventricular cavity overlap as measured by the DICE  
253 similarity coefficient between manual and DL1 contouring in apical 2- and 4-chamber views  
254 for all observers.

255 **Comparison of EF, ESV and EDV values from DL-based contours and manual contours**

256 Table 2 lists the computed ranges of EF, EDV and ESV values from echo images for the whole  
257 population and disaggregated by healthy and pathology groups, when quantified either by the  
258 senior and junior observers or DL1. EF values for most pathological groups were found non  
259 significantly different when measured by A<sub>1</sub> or DL1. DL1 values of EF were also, but to a lesser  
260 extent, consistent with A<sub>2</sub> and B<sub>1</sub>, who both reported higher overall EF values. For EDV and  
261 ESV, there were significant differences between observers, with junior observer A<sub>2</sub> providing  
262 systematically lower EDV and ESV estimates than senior observers A<sub>1</sub> and B<sub>1</sub>.

263 **Table 2. : LV volumes and EF by all observers and modalities.**

	HCM	ISCH	NON-ISCH	VALV	VOL	ALL
<b>End-diastolic volumes [mL]</b>						
<b>Senior A1</b>	120±39	189±69 *	203±64 *	155±58 *	117±24	159±64 ‡
<b>Senior B1</b>	108±33	196±70 *	204±65 *	162±61 *	122±24	163±65 †
<b>Junior A2</b>	72±24 *†‡	157±59 *†‡	154±56 *†‡	128±61 †‡	100±30 †‡	128±57 †‡
<b>DL</b>	91±27 †	165±57 *†‡	165±56 *†‡	122±43 †‡	104±23 †‡	133±53 †‡

<b>CMR</b>	177±53 †‡	233±93 *†‡	264±69 *†‡	199±66 *†‡	150±33 †‡	204±79 †‡
<b>End-systolic volumes [mL]</b>						
<b>Senior A1</b>	52±23 ‡	120±65 *‡	137±57 *‡	69±41 *‡	47±14 ‡	87±58 ‡
<b>Senior B1</b>	35±17 †	113±67 *†	130±63 *†	59±32 †	41±12 †	78±59 †
<b>Junior A2</b>	25±12 †‡	92±56 *†‡	96±46 *†‡	50±29 †‡	38±15 †	64±46 †‡
<b>DL</b>	39±19 †	105±63 *†‡	113±52 *†‡	53±23 *†	38±12 †‡	71±52 †‡
<b>CMR</b>	63±24 ‡	163±97 *†‡	188±70 *†‡	77±45 †‡	57±15 †‡	112±81 †‡
<b>Ejection fractions [%]</b>						
<b>Senior A1</b>	58±12 ‡	40±14 *‡	34±12 *	58±13 ‡	61±6 ‡	50±16 ‡
<b>Senior B1</b>	69±10 †	46±16 *†	38±14 *	64±11 †	67±6 †	56±17 †
<b>Junior A2</b>	66±9 †	45±16 *†	40±11 *	62±10	62±7 ‡	54±15 †‡
<b>DL</b>	57±14	41±17 *‡	33±15 *‡	56±19 *‡	64±7 †‡	50±19 †‡
<b>CMR</b>	65±9	36±16 *†‡	31±13 *‡	63±12 †‡	62±6	50±19 ‡

264 Left ventricle volumes and ejection fraction in the different groups of cardiac conditions as  
 265 measured by manual contouring and DL contouring from 2D echocardiography and semi-  
 266 automated contouring from CMR. See group definitions in **Error! Reference source not**  
 267 **found.** \* indicates p<0.05 vs. the VOL subgroup. † indicates p-values p<0.05 vs. A<sub>1</sub>. ‡  
 268 indicates p-values p<0.01 vs. B<sub>1</sub>.

## 269 Interobserver variability

270 Interobserver agreement (for EF, ESV and EDV) was significantly better for DL1 than manual  
 271 contouring. As illustrated in (Fig 4), this effect was most dominant for LV-EDV where there  
 272 was particularly poor senior-junior (compared to senior-senior) agreement of EDV  
 273 values **Error! Reference source not found.** S1 Table quantifies inter-observer agreement in  
 274 manual vs. DL for the 3 DL algorithms. It shows both DL1 and DL2 reached excellent inter-  
 275 observer agreement (Lin's CCC >0.9 for EDV, ESV and EF) that compared favourably to  
 276 manual inter-observer agreement (Lin's CCC < 0.9 for EDV, ESV and EF)

277 **Fig 4. Lin's concordance correlation plots between observers (top) and for DL1 (bottom,**  
 278 **on the frames quantified by all observers) for EDV in 2D echo.**

## 279 **Agreement of Manual and DL1 measurements with CMR**

280 As shown in Table 2, EF values were consistent between CMR and echo for all observers using  
281 either manual contouring or DL1. Correlation and Bland-Altman plots for EF are shown in (Fig  
282 5). Lin's CCC between echo-EF and CMR-EF improved for A<sub>2</sub> using DL but degraded for  
283 senior observers A<sub>1</sub> and B<sub>1</sub>. The EF bias was reduced with DL1 compared to manual contouring  
284 for junior observer A<sub>2</sub> (-0.1±10.0 vs. 3.7±9.6) and senior observer B<sub>1</sub> (-0.7±13.1 vs 5.9±8.1)  
285 but remained unchanged for A<sub>1</sub> (0.5±11.3 vs -0.5±8.6). Regarding volume measurements, (Fig  
286 6) shows that all echo-based ESV and EDV (using either manual or DL1 for all observers)  
287 values were underestimated w.r.t. CMR. Correlation between EDV by senior observers and  
288 CMR were good (Fig 6) with CCC values over 0.7, higher than the junior cardiologist A<sub>2</sub> (0.45).  
289 This value slightly improved for A<sub>2</sub> using DL (0.51). In comparison, DL1 ESV values were in  
290 better agreement with CMR (CCC > 0.7).

291 **Fig 5. Lin's concordance correlation coefficient and Bland-Altman plot for EF**  
292 **comparing manual and DL1 estimates for each observer in 2D echo to MRI.**

293 **Fig 6. Lin's concordance correlation coefficient plots for LV EDV and LV ESV between**  
294 **CMR and 2D echo for all observers and DL1 applied to the frames of each observer (A<sub>1</sub>**  
295 **[senior], A<sub>2</sub> [junior] and B<sub>1</sub> [senior]).**

## 296 **Comparison to other DL methods**

297 (Fig 7) compares DL1, DL2 and DL3 agreement measurements on all frames selected by  
298 observers with CMR. DL1 and DL2 exhibited similar Lin's CCC for volume evaluation, while  
299 DL3 performed less well (0.51 resp. 0.49 vs. 0.22 for EDV, and 0.73 resp. 0.74 vs. 0.53 for  
300 ESV). EF values were not valid in 12.7% of the cases for DL3 vs. 0.78% for DL1 and 1.56%  
301 for DL2. All three DL methods showed acceptable Pearson correlation (0.86 for DL1, 0.72 for  
302 DL2, and 0.57 for DL3), but only DL1 and DL2 showed acceptable agreement with CMR. DL1  
303 achieved higher agreement than DL2 due to a reduced bias and SD, thus contrasting with their  
304 more homogeneous EDV / ESV findings.

305 **Fig 7. Agreement between CMR and 2D echo for all observers and the three DL**  
306 **techniques applied to the frames of each observer (A1 [senior], A2 [junior] and B1 [senior]).**  
307 DL1 is the network proposed in this paper, DL2 is the e-Net architecture from (13), DL3 is the  
308 network proposed in (18).

### 309 **Impact on population stratification**

310 When CMR and echo LVEF were classified into three categories (<40, 40-50 and >50% EF),  
311 Cohen's kappa agreement to CMR EF labels was similar between manual and DL1 contouring  
312 ( $0.65\pm 0.11$  vs  $0.61\pm 0.14$ ) compared to  $0.48\pm 0.11$  for DL2 and  $0.29\pm 0.11$  for DL3. When  
313 confronting echo vs. CMR EF labels for A<sub>1</sub>'s frames, manual contouring misclassified 22  
314 patients, and DL contouring misclassified 16 patients. However, for A<sub>2</sub> and B<sub>1</sub>, the opposite  
315 situation was observed: there were 33 misclassified subjects using manual contouring for A<sub>2</sub>  
316 vs. 27 using DL1 (35 vs. 19 for B<sub>1</sub>).

## 317 **Discussion**

318 The principal findings of our study can be summarized as follows.

319 First, our DL algorithm (DL1) trained echo contouring performed well in an unrelated  
320 population of patients despite a balanced dataset in terms of image quality. It generalized well  
321 to data obtained in another centre (B) representing 80% of cases.

322 Second, DL1 reduced interobserver agreement relative to manual contouring for LV-EF, EDV  
323 and ESV, and in particular between junior and senior observers. This was confirmed for the  
324 other two compared DL algorithms, tending to suggest that DL can be instrumental in  
325 increasing the interobserver reproducibility of 2D echo-based EF values, if taken as an initial  
326 contour before potential manual edits, when required in more challenging cases.

327 Third, this study demonstrated that DL1 compared favourably to manual contouring in terms  
328 of EF accuracy when taking CMR as reference (Fig 6). EF bias was brought to almost zero  
329 when segmenting the same image frames as the 3 observers with DL1, which reduced the bias



330 of one junior and one senior observer. Yet biases in LV volumes remained present and DL1 did  
331 not correct the underestimation of 2D echo-based volumes, compared to CMR. Such  
332 underestimation is believed to result in part from foreshortening of acquired 2D, this bias is  
333 known to be reduced by 3D echo. It may also result from differences in detection of trabeculated  
334 myocardium. Accordingly, DL algorithms trained with both echo and CMR data might allow  
335 learning some of this systematic bias. Adding more pathological groups to the training database  
336 could potentially improve EF biases for different disease groups.

337 When comparing our DL results with previously trained network, we found (Fig 7) better  
338 agreement with CMR EF and reduced bias than DL2 and DL3. However, DL2 appeared as a  
339 clear contender and showed excellent inter-observer agreement and a good correlation with  
340 CMR. It can be argued the comparison done in this paper is unfair, as DL1 was trained on data  
341 from clinical centre A, all performed on Philips echocardiographic devices, with manual  
342 contouring from observer A1. DL2 was trained on ground truths segmentations from other  
343 observers and on GE data. Therefore, the comparison presented here should be taken as a direct  
344 test of generatability of an echocardiographic DL segmentation algorithm (DL2) without  
345 applying any transfer learning to another constructor and possibly with other contouring  
346 conventions. Our results illustrate the need to learn models that generalize well across vendors  
347 and clinical centres, possibly through federated learning. DL3 was applied similarly without  
348 any and adaptation and performed poorly on our data. As for DL2, this probably reflects  
349 discrepancies between the training data of DL1 and DL3 and calls for further adaptation of the  
350 DL3 network to DL1 training data that are beyond the scope of this paper.

351 Finally, using CMR-based EF reference values, we could evaluate the potential impact that an  
352 echo- and DL-based EF computation would have on the stratification of a Heart Failure  
353 population. We found a similar ( $A_1$ ) or improved ( $A_2$ ,  $B_1$ ) agreement between echo  
354 measurements vs. CMR using the DL algorithm over manual contouring. This preliminary

355 finding should be confirmed on a population of HF patients with preserved and reduced EF to  
356 determine whether or not the added value of DL vs. manual contouring is confirmed.

357 Several previous studies did report agreement and correlation values of automated and manual  
358 EF values. The AutoEF algorithm was evaluated in large studies (>200 subjects (21)) but the  
359 comparison to CMR could only be performed for a subset (~20) of the population. In (22), the  
360 AutoEF results were edited when deemed necessary by both senior and novice observers, which  
361 represents a potential bias when comparing manual and automated contouring solutions. Other  
362 commercial algorithms (23) were also assessed against manual contouring but often without  
363 involving another modality as reference. As DL-based segmentation solutions are emerging in  
364 echocardiography (13,24), they need to be benchmarked not only for accuracy against manual  
365 observers but also against other imaging modalities, and more specifically against CMR as it  
366 stands as a gold standard modality for LV EF assessment like CMR.

367 Most of EF validation studies that took CMR as reference were comparing 2D echo to 3D echo,  
368 and demonstrated a higher accuracy on EDV and ESV measurements (24), as well as lower  
369 intra- and interobservers variability (25) and higher performance for some pathologies such as  
370 HCM (26). Nonetheless, the spatiotemporal resolution of 3D echo, which is inherently lower  
371 than 2D imaging, can be challenging with larger chambers. In addition, 3D echo remains a  
372 premium imaging modality, not as widespread as 2D echo. Improving the echocardiographic  
373 workflow involves automating time-consuming tasks for 2D echo images as well as 3D echo.

374 However, the processing of 2D echo is still mostly manual, unlike 3D echo, for which advanced  
375 model-based (editable) segmentation algorithms are available (25,26). This situation called for  
376 a thorough evaluation of a modern automated segmentation on 2D echo, validating it with  
377 another 3D reference modality.

378 By contributing an open validation dataset, together with the bi-plane Simpson code, this paper  
379 contributes a reproducible evaluation framework, against which other DL methods can be  
380 benchmarked.

### 381 **Clinical implications**

382 We argue that the framework described here could help exploit the full potential of deep  
383 learning for echocardiographic applications

384 - simplifying LVEF and volume calculations to allow for multi-cycle or real-time assessment.

385 - Improved longitudinal follow-up of chronic patients due to good overall agreement with CMR  
386 and reduced inter-observer variability.

387 - Improved management strategies due to the accuracy of the LVEF category classification.

### 388 **Study Limitations**

389 In this paper, the observers, not to interfere with clinical practice, were free to choose the cycles  
390 and frames on which they quantified EDV and ESV volumes. Therefore, we could not compute  
391 local contour differences between observers. Such a local analysis could have revealed regions  
392 of higher variability or systematic interobserver differences. Similarly, we could not study if  
393 the DL segmentation represents a good consensus by comparing its contour to the observers'  
394 contours. A further automatization could include a separate pre-processing DL network  
395 automatically selecting the ES and ED frames. This was left as future work and likely requires  
396 a separate evaluation.

397 We limited our comparisons to EDV, ESV and EF values, as they appeared as a priority, being  
398 clinical indices used routinely. Yet this approach probably better reflected clinical practice,  
399 where there is also intersubject variability in selection of frames. The clinical centres compared  
400 in this study have similar protocols in terms of echocardiography and used the same equipment.  
401 The algorithm might behave less accurately on other echocardiographic systems or image  
402 acquisition protocols. Extending the analysis of this paper to other clinical centres could further

403 span differences across countries in terms of conventions for defining the endocardial contour,  
404 in terms of expertise (e.g. junior sonographer vs senior cardiologist), or in terms of time  
405 constraints for the echocardiographic exam. Also, we did not cover in this study reproducibility  
406 issues stemming from the acquisition (e.g. probe orientation) that can induce foreshortening.  
407 Finally, although CMR is widely accepted as reference modality for the validation of echo-  
408 based measurements, measurements performed on short axis slices only could underestimate  
409 the long axis contribution of LV motion (27).

## 410 **Conclusions**

411 In this paper, we compared manual and DL automated contouring from 2D echocardiographic  
412 images with respect to CMR, taking the latter as reference for the computation of EF, ESV and  
413 EDV values. We demonstrated the value of a DL-based automated contouring of AP2 and AP4  
414 images to reduce and homogenize the biases in EF with respect to CMR. This study also  
415 confirmed important biases in EDV and ESV 2D echo-based values, for automated and manual  
416 contouring, that nonetheless get compensated when computing EF, reaching a practically null  
417 bias between CMR and echo-based EF values.

## 418 **Funding**

419 This research received no specific grant from any funding agency in the public, commercial,  
420 or not-for-profit sectors.

## 421 **References**

422 1. Bonow RO., Lakatos E., Maron BJ., Epstein SE. Serial long-term assessment of the  
423 natural history of asymptomatic patients with chronic aortic regurgitation and normal left  
424 ventricular systolic function. *Circulation* 1991;84(4):1625–35.

- 425 2. Enriquez-Sarano M., Tajik AJ., Schaff H V., Orszulak TA., Bailey KR., Frye RL.  
426 Echocardiographic prediction of survival after surgical correction of organic mitral  
427 regurgitation. *Circulation* 1994;90(2):830–7.
- 428 3. Bohbot Y., de Meester de Ravenstein C., Chadha G., et al. Relationship Between Left  
429 Ventricular Ejection Fraction and Mortality in Asymptomatic and Minimally Symptomatic  
430 Patients With Severe Aortic Stenosis. *JACC Cardiovasc Imaging* 2019;12(1):38–48.
- 431 4. Rouleau JL., Talajic M., Sussex B., et al. Myocardial infarction patients in the 1990s—  
432 their risk factors, stratification and survival in Canada: The Canadian assessment of myocardial  
433 infarction (CAMI) study. *J Am Coll Cardiol* 1996;27(5):1119–27.
- 434 5. Buxton AE., Lee KL., DiCarlo L., et al. Electrophysiologic Testing to Identify Patients  
435 with Coronary Artery Disease Who Are at Risk for Sudden Death. *N Engl J Med*  
436 2000;342(26):1937–45.
- 437 6. Rihal CS., Nishimura RA., Hatle LK., Bailey KR., Tajik AJ. Systolic and diastolic  
438 dysfunction in patients with clinical diagnosis of dilated cardiomyopathy. Relation to symptoms  
439 and prognosis. *Circulation* 1994;90(6):2772–9.
- 440 7. Ponikowski P., Voors AA., Anker SD., et al. 2016 ESC Guidelines for the diagnosis and  
441 treatment of acute and chronic heart failure. *Eur J Heart Fail* 2016;18(8):891–975
- 442 8. Knuuti J., Wijns W., Saraste A., et al. 2019 ESC Guidelines for the diagnosis and  
443 management of chronic coronary syndromes. *Eur Heart J* 2020;41(3):407–77.
- 444 9. Baumgartner H., Falk V., Bax JJ., et al. 2017 ESC/EACTS Guidelines for the  
445 management of valvular heart disease. *Eur Heart J* 2017;38(36):2739–91
- 446 10. Otterstad JE. Measuring left ventricular volume and ejection fraction with the biplane  
447 Simpson’s method. *Heart* 2002;88(6):559–60.

- 448 11. Otterstad JE., Froeland G., St John Sutton M., Holme I. Accuracy and reproducibility  
449 of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and  
450 function. *Eur Heart J* 1997;18(3):507–13.
- 451 12. Weese J., Groth A., Nickisch H., et al. Generating anatomical models of the heart and  
452 the aorta from medical images for personalized physiological simulations. *Med Biol Eng*  
453 *Comput* 2013;51(11).
- 454 13. Leclerc S., Smistad E., Pedrosa J., et al. Deep Learning for Segmentation Using an Open  
455 Large-Scale Dataset in 2D Echocardiography. *IEEE Trans Med Imaging* 2019;38(9):2198–210.
- 456 14. Howard JP., Stowell CC., Cole GD., et al. Automated Left Ventricular Dimension  
457 Assessment Using Artificial Intelligence Developed and Validated by a UK-Wide  
458 Collaborative. *Circ Cardiovasc Imaging* 2021;14(5)
- 459 15. Mitchell C., Rahko PS., Blauwet LA., et al. Guidelines for Performing a Comprehensive  
460 Transthoracic Echocardiographic Examination in Adults: Recommendations from the  
461 American Society of Echocardiography. *J Am Soc Echocardiogr* 2019;32(1):1–64.
- 462 16. Folland ED., Parisi AF., Moynihan PF., Jones DR., Feldman CL., Tow DE. Assessment  
463 of left ventricular ejection fraction and volumes by real-time, two-dimensional  
464 echocardiography. A comparison of cineangiographic and radionuclide techniques. *Circulation*  
465 1979;60(4):760–6.
- 466 17. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical  
467 Image Segmentation. In: Navab N, Hornegger J, Wells W, and Frangi AF, editors. *Medical*  
468 *Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, Cham;  
469 2015. p. 234–41.
- 470 18. Zhang J., Gajjala S., Agrawal P., et al. Fully Automated Echocardiogram Interpretation  
471 in Clinical Practice. *Circulation* 2018;138(16):1623–35.

- 472 19. Heiberg E., Ugander M., Engblom H., et al. Automated quantification of myocardial  
473 infarction from MR images by accounting for partial volume effects: animal, phantom, and  
474 human study. *Radiology* 2008;246(2):581–8..
- 475 20. Ponikowski P., Voors AA., Anker SD., et al. 2016 ESC Guidelines for the diagnosis and  
476 treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of  
477 acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with  
478 the special contribution. *Eur J Heart Fail* 2016;18(8):891–975.
- 479 21. Rahmouni HW., Ky B., Plappert T., et al. Clinical utility of automated assessment of  
480 left ventricular ejection fraction using artificial intelligence-assisted border detection. *Am Heart*  
481 *J* 2008;155(3):562–70.
- 482 22. Maret E., Brudin L., Lindstrom L., Nylander E., Ohlsson JL., Engvall JE. Computer-  
483 assisted determination of left ventricular endocardial borders reduces variability in the  
484 echocardiographic assessment of ejection fraction. *Cardiovasc Ultrasound* 2008;6:45–54.
- 485 23. Knackstedt C., Bekkers SCAM., Schummers G., et al. Fully Automated Versus  
486 Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain the FAST-EFs  
487 Multicenter Study. *J Am Coll Cardiol* 2015;66(13):1456–66.
- 488 24. Silva JF., Silva JM., Guerra A., Matos S., Costa C. Ejection Fraction Classification in  
489 Transthoracic Echocardiography Using a Deep Learning Approach. 2018 IEEE 31st  
490 International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2018. p. 123–  
491 8.
- 492 25. Tamborini G., Piazzese C., Lang RM., et al. Feasibility and Accuracy of Automated  
493 Software for Transthoracic Three-Dimensional Left Ventricular Volume and Function  
494 Analysis: Comparisons with Two-Dimensional Echocardiography, Three-Dimensional  
495 Transthoracic Manual Method, and Cardiac Magnetic Resona. *J Am Soc Echocardiogr*  
496 2017;30(11):1049–58.

497 26. Jacobs LD., Salgo IS., Goonewardena S., et al. Rapid online quantification of left  
498 ventricular volume from real-time three-dimensional echocardiographic data. *Eur Heart J*  
499 2006;27(4):460–8.

500 27. Tufvesson J., Hedstrom E., Steding-Ehrenborg K., Carlsson M., Arheden H., Heiberg  
501 E. Validation and Development of a New Automatic Algorithm for Time-Resolved  
502 Segmentation of the Left Ventricle in Magnetic Resonance Imaging. *Biomed Res Int*  
503 2015;2015:970357.

## 504 **Supporting information**

505 **S1 Fig. Overlay of DL segmentations (yellow mask) and manual contouring (red dotted**  
506 **contour) for 6 outliers.** Frame and manual segmentation performed by observer A<sub>1</sub> (top row),  
507 A<sub>2</sub> (middle row) and B<sub>1</sub> (bottom row). Outliers involved subjects with dilated non-ischaemic  
508 cardiomyopathy (NON-ISCH), mitral regurgitation or aortic stenosis (VALV) and  
509 hypertrophic cardiomyopathy (HCM).

510  
511 **S2 Fig. Lin's concordance correlation plots between CMR and 2D echo for the DL2 and**  
512 **DL3 algorithms applied to the frames of each observer (A<sub>1</sub> [senior], A<sub>2</sub> [junior] and B<sub>1</sub>**  
513 **[senior]) for LV EDV and LV ESV.**

514  
515 **S1 Table. Agreement between observers in echo for ESV, EDV and EF with manual and**  
516 **DL contouring using Lin's concordance correlation coefficient**

517



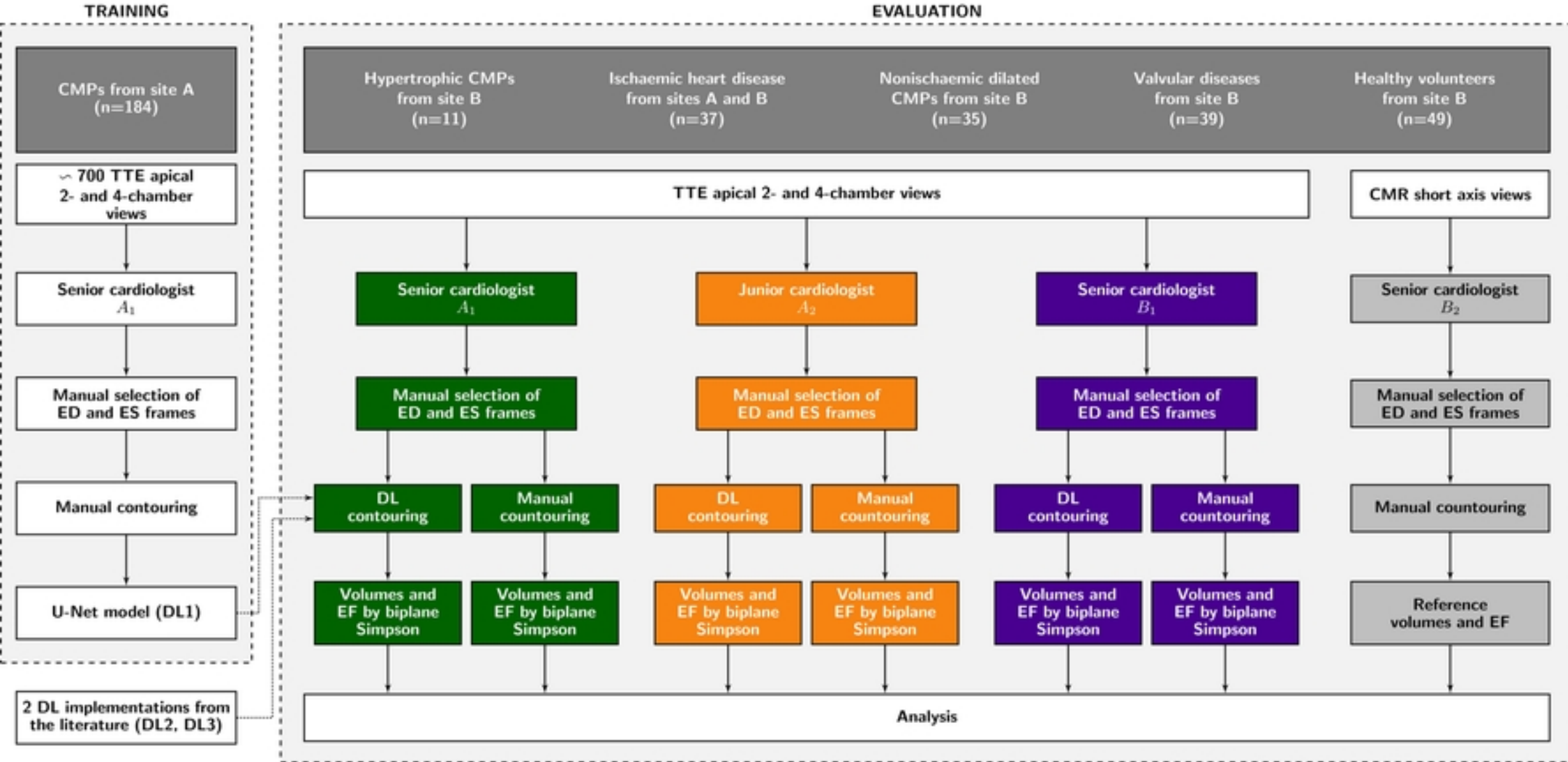


Fig 1.

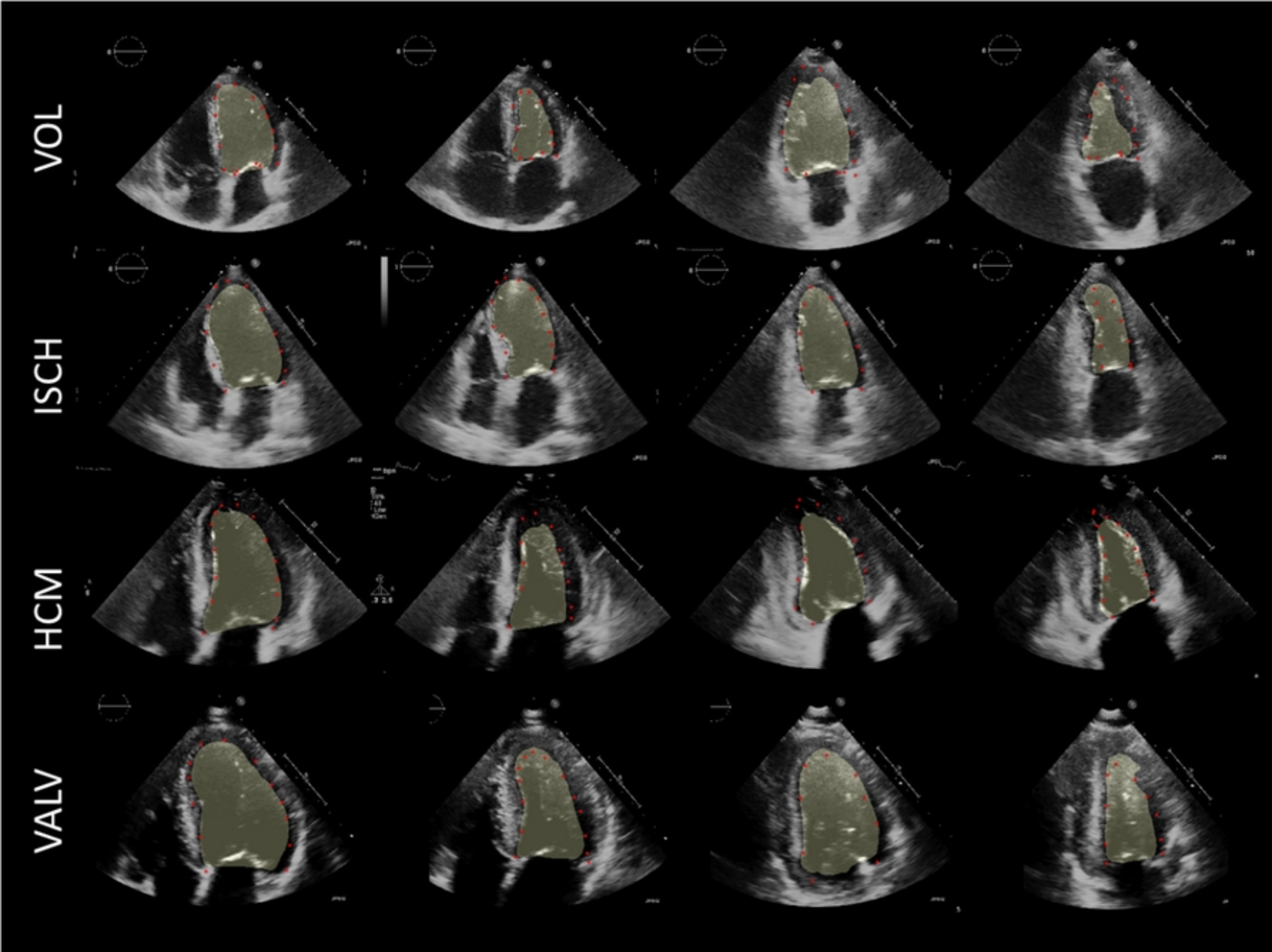
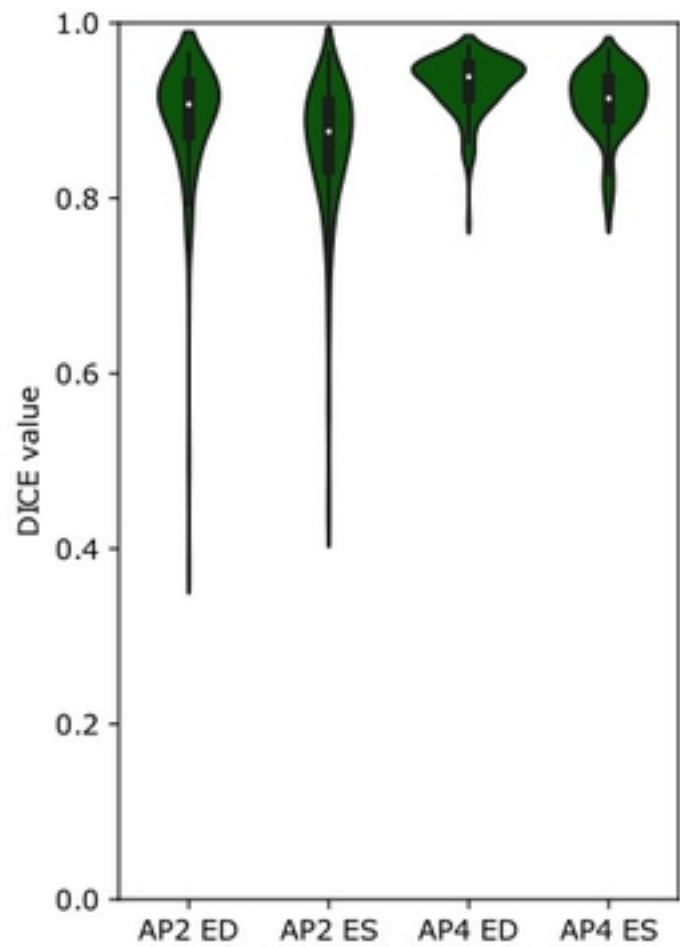
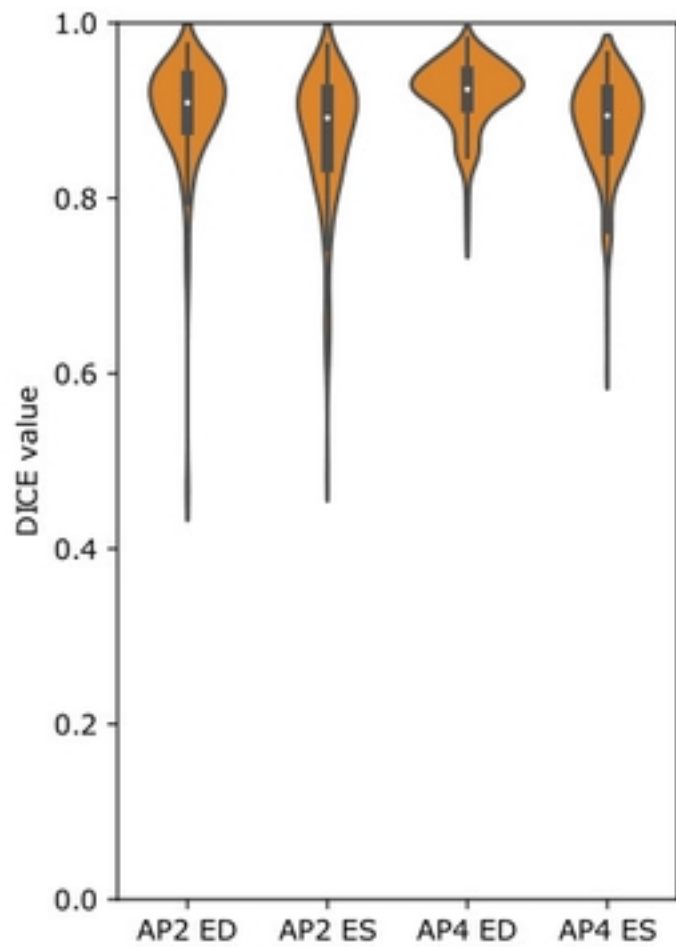


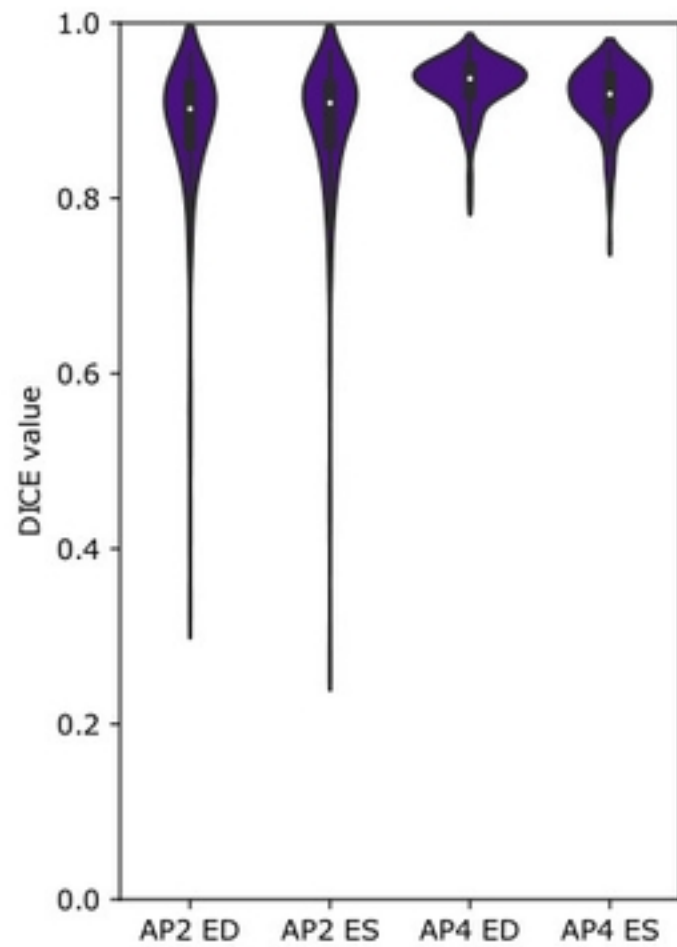
Fig 2.



(a) Observer A1 (Senior)



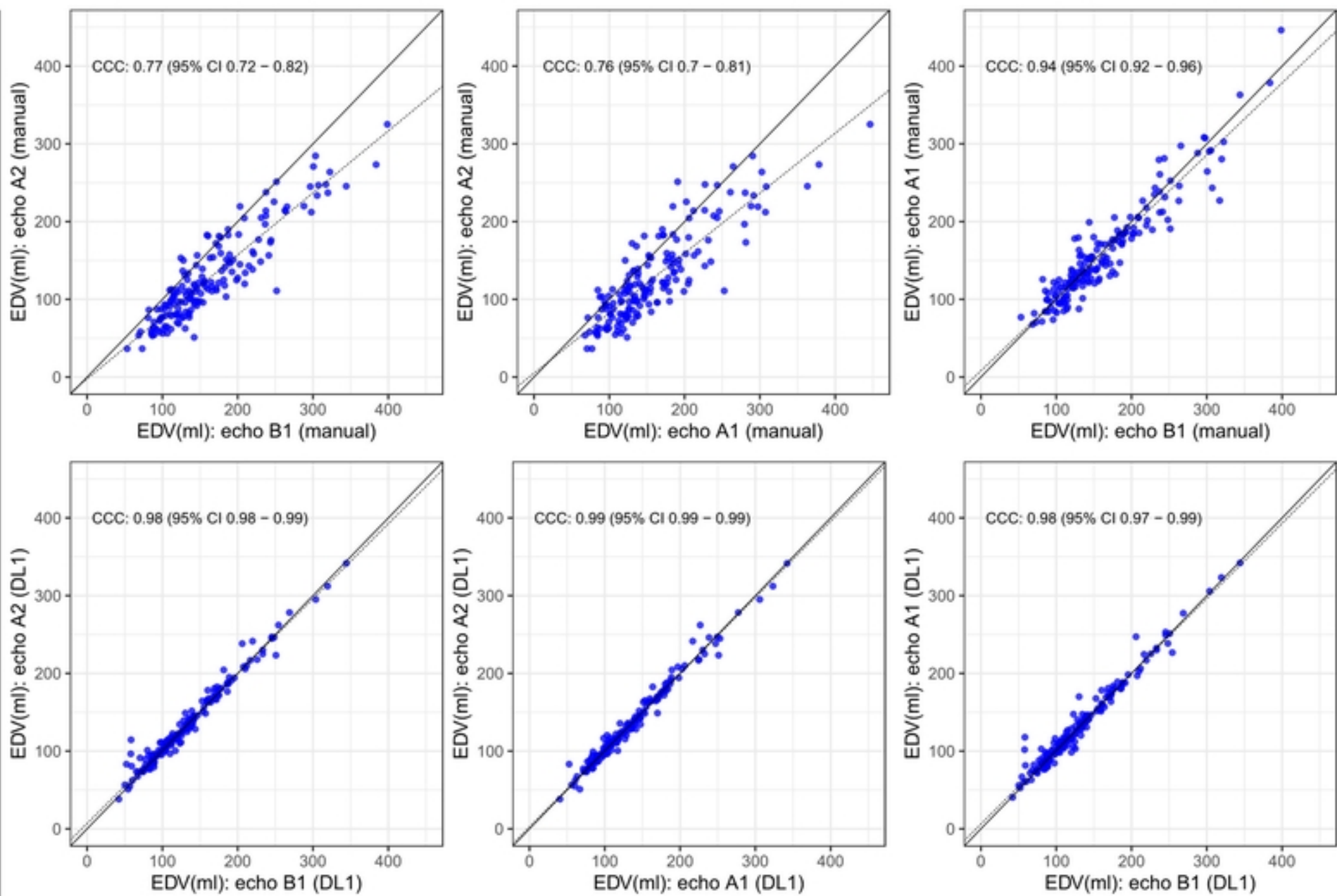
(b) Observer A2 (Junior)



(c) Observer B1 (Senior)

Fig 3.

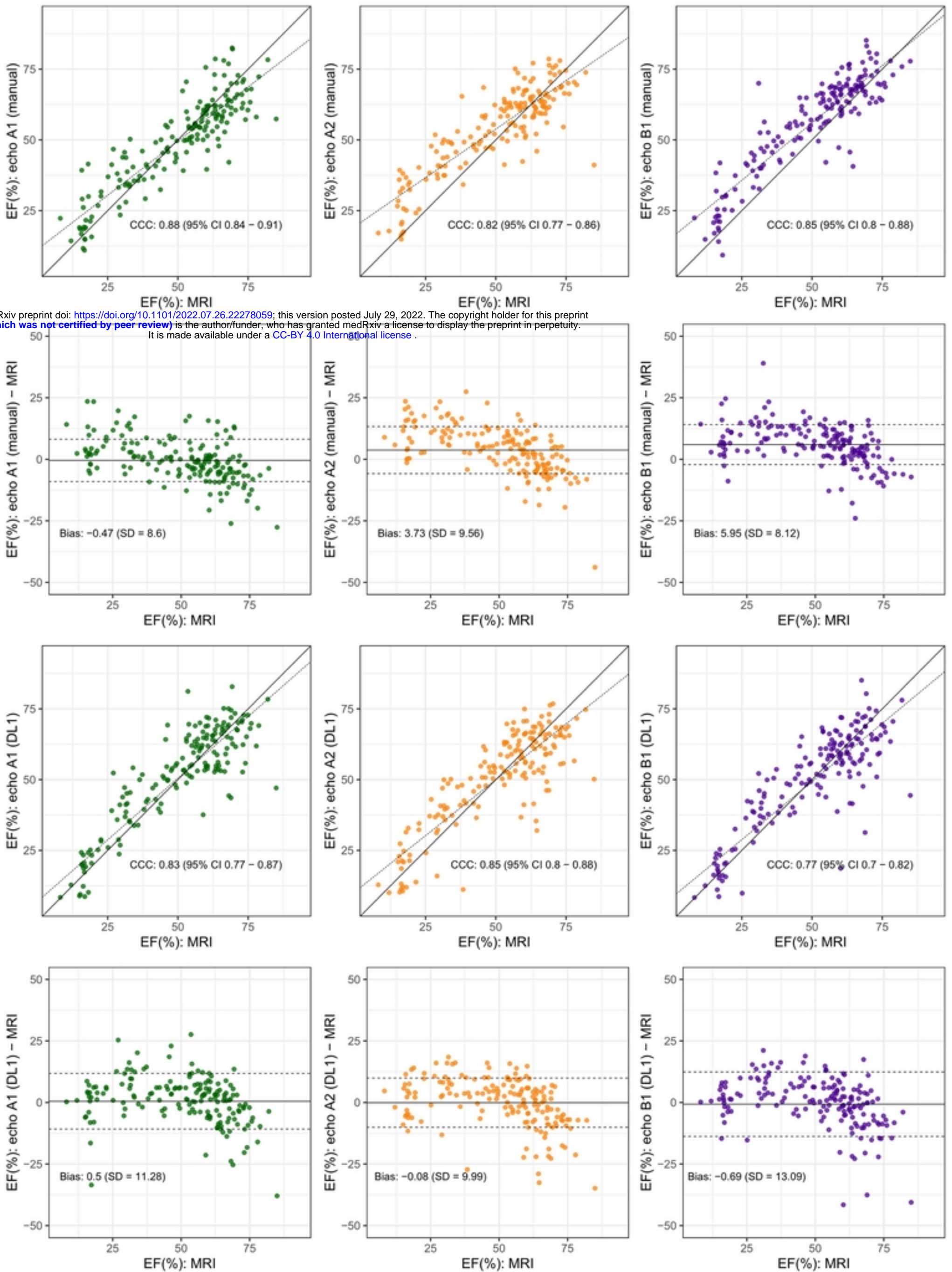
**E  
D  
V**



**Manual**

**DL1**

Fig 4.

**A1****A2****B1**

Rxiv preprint doi: <https://doi.org/10.1101/2022.07.26.22278059>; this version posted July 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

T M

Manual

DL1

Fig 5.

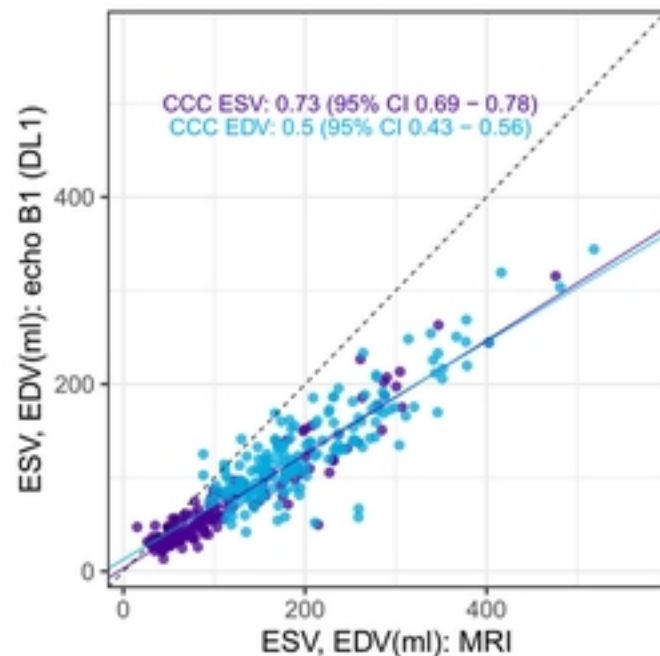
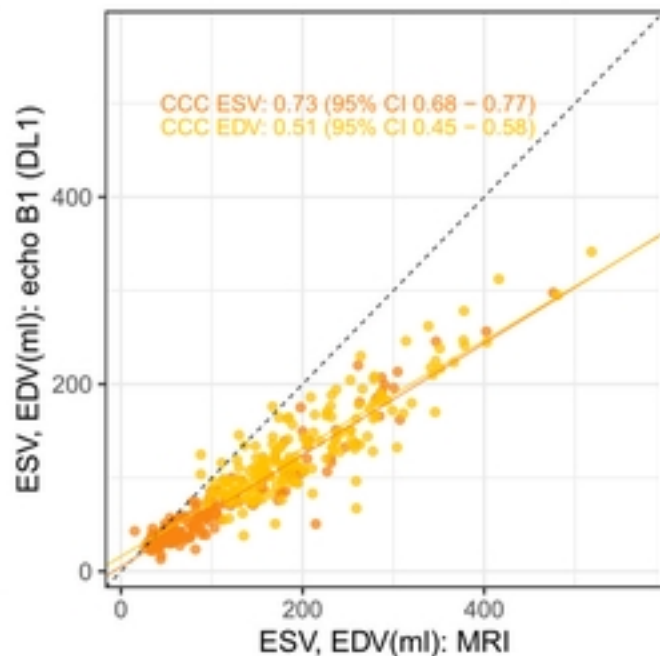
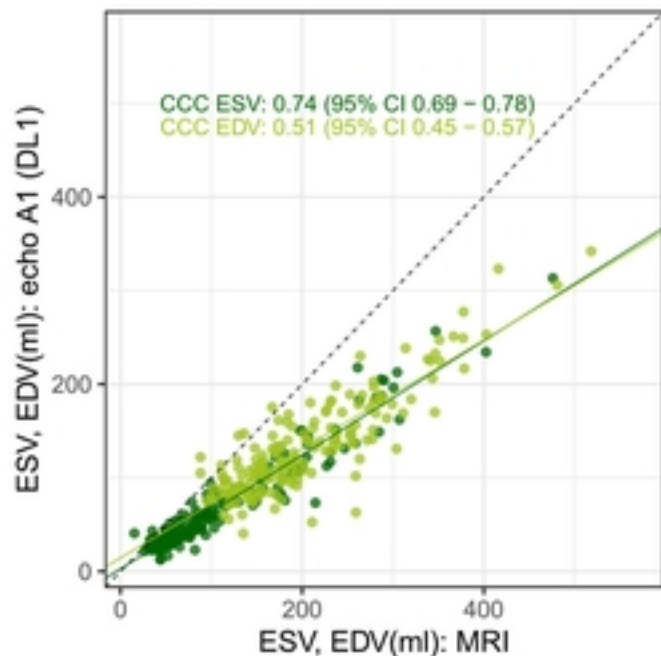
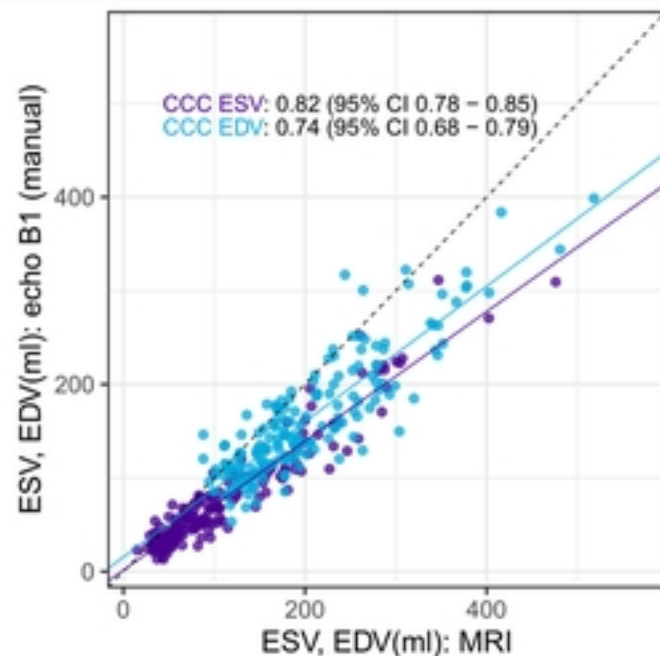
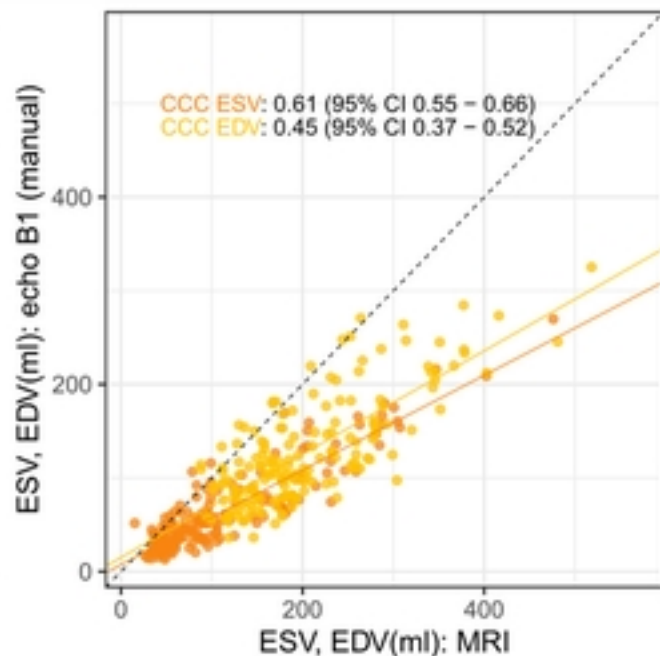
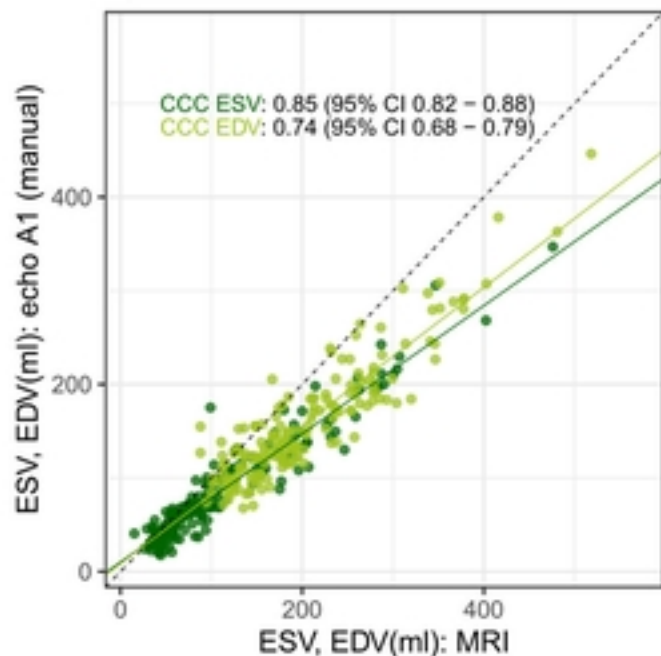
**A1****A2****B1****ESV & EDV****Manual****DL1**

Fig 6.

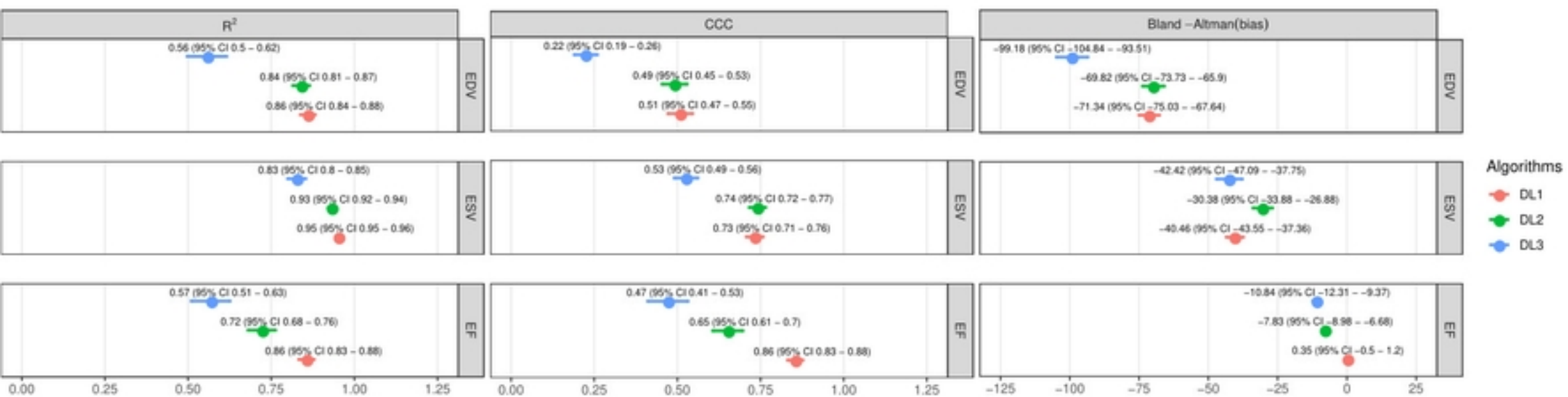


Fig 7.