Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Clinical Performance and Role of Expert Supervision of Deep Learning for Cardiac Ventricular Volumetry: A Validation Study

Tara A. Retson, MD, PhD • Evan M. Masutani, BS • Daniel Golden, PhD • Albert Hsiao, MD, PhD

From the Department of Radiology, Altman Clinical and Translational Research Institute, University of California, San Diego, 9452 Medical Center Dr, 4th Floor, La Jolla, CA 92037 (T.A.R., A.H.); Department of Bioengineering, University of California San Diego School of Medicine, La Jolla, Calif (E.M.M.); and Arterys, San Francisco, Calif (D.G.). Received April 24, 2019; revision requested June 11; revision received February 21, 2020; accepted March 27. **Address correspondence to** T.A.R. (e-mail: *tretson@ucs.edu*).

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

Radiology: Artificial Intelligence 2020; 2(4):e190064 • https://doi.org/10.1148/ryai.2020190064 • Content codes: CA IN

Purpose: To evaluate the performance of a deep learning (DL) algorithm for clinical measurement of right and left ventricular volume and function across cardiac MR images obtained for a range of clinical indications and pathologies.

Materials and Methods: A retrospective, Health Insurance Portability and Accountability Act–compliant study was conducted using the first 200 noncongenital clinical cardiac MRI examinations from June 2015 to June 2017 for which volumetry was available. Images were analyzed using commercially available software for automated DL-based and manual contouring of biventricular volumes. Fully automated measurements were compared using Pearson correlations, relative volume errors, and Bland-Altman analyses. Manual, automated, and expert revised contours for 50 MR images were examined by comparing regional Dice coefficients at the base, midventricle, and apex to further analyze the contour quality.

Results: Fully automated and manual left ventricular volumes were strongly correlated for end-systolic volume (ESV: Pearson $r = 0.99$, $P < .001$), end-diastolic volume (EDV: $r = 0.97$, $P < .001$), and ejection fraction (EF: $r = 0.94$, $P < .001$). Right ventricular measurements were also correlated for ESV ($r = 0.93$, $P < .001$), EDV ($r = 0.92$, $P < .001$), and EF ($r = 0.73$, $P < .001$). Visual inspection of segmentation quality showed most errors (73%) occurred at the cardiac base. Mean Dice coefficients between manual, automated, and expert revised contours ranged from 0.92 to 0.95, with greatest variance at the base and apex.

Conclusion: Fully automated ventricular segmentation by the tested algorithm provides contours and ventricular volumes that could be used to aid expert segmentation, but can benefit from expert supervision, particularly to resolve errors at the basal and apical slices.

Supplemental material is available for this article.

Cardiac MRI is a powerful, noninvasive modality used for the diagnosis and management of a wide variety of cardiovascular diseases and is the clinical reference standard for quantification of cardiac volumetry and function (1–4). Postprocessing and analysis of cardiac MR images can be challenging because it requires considerable time and expertise. To obtain volumetric measurements, physicians must manually draw contours outlining the endocardium and epicardium for 10 to 15 short-axis slices (3,5,6). Even experienced physicians may require 20 to 30 minutes per examination to perform these tasks (7). In addition, heterogeneity in the approach to ventricular segmentation can lead to interreader variability (6,8–10). In light of these considerations, several groups have sought to reduce the time burden of manual quantification and decrease subjectivity (11) by using machine learning approaches (12–14). Traditional machine learning algorithms have emphasized specific features, such as edge detection or probabilistic atlases to reduce processing times, but are prone to artifact and have limited generalizability (11,15–17). New machine learning and deep learning (DL) algorithms are showing promise for automating or supporting technical tasks, such as selection of imaging planes and myocardial inversion time (18), to aid radiologists in image assessment.

DL applications are being developed to overcome the limitations of traditional machine learning for cardiac segmentation and volumetry (9,19–22). Multiple groups have used DL approaches to compete in machine learning challenges and have achieved excellent performance (9,23,24). However, because most common training sets are limited in scope and may not reflect the range of pathologies seen in clinical practice (25–28), the performance of these algorithms in a clinical setting requires further investigation.

In this study, we aimed to test the hypothesis that a DL–based algorithm is capable of biventricular segmentation and volumetry on clinical cardiac MRI data. We evaluated a two-dimensional U-Net convolutional neural network on data from 200 clinical cardiac MRI examinations by comparing its performance with measurements made by experienced physicians in the course of clinical practice. We then explored the similarity of segmentations to identify areas that could benefit from further improvement and continued physician supervision. We found that the DL algorithm tested was able to segment the right and left ventricles, suggesting that this

Abbreviations

DL = deep learning, EDV = end-diastolic volume, ED = end diastole, EF = ejection fraction, ES = end systole, ESV = end-systolic volume, LV = left ventricular, RV = right ventricular

Summary

Performance of a deep learning–based automated ventricular segmentation algorithm is similar to expert segmentation and may provide streamlined interpretation of cardiac MR images.

Key Points

- The tested algorithm is capable of segmenting right and left ventricular contours that are strongly correlated with fully manual measurements by radiologists, which may help streamline quantitative interpretation.
- Algorithm performance is optimal at the midventricular slices but requires greater supervision for the right ventricle, cardiac base and apex, and end-systolic phases.

application could be used to streamline image interpretation by a practicing radiologist.

Materials and Methods

Study Design

With Health Insurance Portability and Accountability Act compliance and institutional review board waiver of informed consent, cardiac MRI examinations were identified that were performed at our institution between June 2015 and June 2017. The first 200 examinations were retrospectively identified in which cardiac volumes were available as part of the original clinical examination. Patients with complex congenital heart disease were not included. No other exclusions were made based on clinical study purpose or nature of pathology, with clinical indications listed in Table 1. The average patient age was 55 years (range, 18–88 years); 59% were men and 41% were women.

Table 1: Study Indications Obtained from the Clinical Record

Parameter	Value
Patient demographics	
No. of men	118 (59.0)
No. of women	82 (41.0)
Average age (y)	55 (18–88)*
Study indications	
Cardiomyopathy (including hypertrophic)	44 (22.0)
Arrhythmia or syncope	25 (12.5)
Evaluation of mass, thrombus, or abnormal echo	23 (11.5)
Myocardial viability	23 (11.5)
Heart failure	22 (11.0)
Myocarditis or sarcoid	16 (8.0)
Evaluation for a procedure	9 (4.5)
Congenital	5 (2.5)
Other	33 (16.5)

Note.—Unless otherwise indicated, data are numbers of patients with percentages in parentheses. A total of 200 patients were included within this study.
* Age range is in parentheses.

MRI Parameters

The examinations were performed for a range of clinical indications with a single Signa HDxt 1.5-T scanner (GE Medical Systems, Wis). Imaging protocols included a cine cardiac-gated steady-state free-precession short-axis stack for cardiac volumetry, performed as part of the clinical examination. Imaging parameters included an average echo time of 1.8 msec, average repetition time of 4.1 msec, and average temporal resolution of 57 msec, with flip angles at 55° and slice thicknesses at 8 mm. Examination indications varied and included new onset heart failure, myocardial viability, hypertrophic cardiomyopathy, and congenital disease (see Table 1 for study indications drawn from the clinical record).

Image Analysis

Manual measurements of biventricular volume were independently performed at the time of clinical examinations by one of five board-certified radiologists (including A.H.) in our institution's cardiothoracic imaging division using Cvi42 v5.3.8 (Circle Cardiovascular Imaging, Calgary, Alberta, Canada). Measurements were obtained by manually tracing the left ventricular (LV) epicardium and endocardium and right ventricular (RV) endocardium borders for 10–15 slices through the cardiac short axis at end systole (ES) and end diastole (ED). DL analysis was performed retrospectively on the same images using a commercially available and Food and Drug Administration–cleared two-dimensional U-Net–based convolutional neural network (Cardio DL 2.3; Arterys, San Francisco, Calif) (19). The study presented here represents an independent test of this algorithm because the Cardio DL 2.3 algorithm was not trained with any data from our institution.

The original clinical segmentations were available for 50 of the studies performed between June 2015 and May 2016 as a result of technical limitations in data available for export. To examine contour quality in more granular detail, a subanalysis was performed on these cases. In-house software was developed in Python for extraction of contour coordinates from their original saved files, and one case was removed for misregistration. Three sets of contours were compared in this analysis, including *(a)* the original manual contours performed at the time of clinical examination by a radiologist, *(b)* fully automated contours from DL obtained retrospectively, and *(c)* an additional set of expert revised contours in which DL contours were edited by a board-certified cardiothoracic radiologist with 10 years of experience (A,H.) using Cardio DL 2.3. A minimum of 18 months elapsed between clinical quantification and the blinded secondary analysis. To minimize potential conflicts of interest, coauthors with financial interest in either software product were not involved in case selection or data analysis, and no direct compensation or software was provided to any author for his or her work on this study.

Table 2: Volumetry Comparisons across Groups

Volume	Correlation (*r*)	Mean Difference (mL or %)	2 SD	Relative Volume Error (%)
	Volumes Measured as Part of the Clinical Examination Compared with Fully Automated (All Cases)			
LV EDV	0.97	24.3	31.1	13.6
LV ESV	0.99	10.9	25.5	−9.9
LV EF	0.94	1.7%	11.2	3.1
RV EDV	0.92	19.6	45.4	10.6
RV ESV	0.93	1.4	32.6	5.2
RV EF	0.73	8.0%	19.1	14.1
	Manual Compared with Expert Revised Subanalysis			
LV EDV	0.98	15	27.1	−8.2
LV ESV	0.97	11.7	28.9	−12.8
RV EDV	0.96	1.4	32	−0.4
RV ESV	0.9	−0.9	28.6	3.8
Overall average	0.97	7.1	32.4	−4.7
	Manual Compared with Automated Subanalysis			
LV EDV	0.97	29.8	33.1	−16.9
LV ESV	0.97	11.7	28.8	−11
RV EDV	0.95	17.2	39.9	−9.8
RV ESV	0.89	−4.9	29	12.3
Overall average	0.96	14.4	41.4	−7.3
	Expert Revised Compared with Automated Subanalysis			
LV EDV	0.98	14.9	24	−9.4
LV ESV	0.97	−0.2	26.7	4.4
RV EDV	0.97	16.8	31.5	−10.3
RV ESV	0.91	−3	27.1	7.1
Overall average	0.97	7.2	32.6	−2.1

Note.—All significance values for correlations are $P < .001$. EDV = end-diastolic volume, EF = ejection fraction, ESV = end-systolic volume, LV = left ventricle, RV = right ventricle, SD = standard deviation.

Comparative Analyses

Manual measurements of RV and LV volume were considered "ground truth" and formed the basis for comparison. Comparisons were made for end-diastolic volume (EDV) and end-systolic volume (ESV), as well as ejection fractions (EFs). ES was visually determined as the phase with the smallest midventricular size, often correlating with the phase just before valve opening. Statistical analysis included Pearson correlation, relative volume error, and bias (shown in Bland-Altman plots). In addition to quantitative comparisons, we performed a visual comparison of contours for each of the 200 cases at both ED and ES and tabulated errant inclusion or exclusion of apical or basal slices, nonanatomic shapes, and nonanatomic location of contours. The location (base, mid, or apex) and the cardiac phase (ES or ED) of these errors were also noted. It was possible for an errant contour to be counted as multiple error types. For example, a contour placed beyond the basal extent of a ventricle and shaped discordantly with natural contours would be tabulated as both an errant contour inclusion and a nonanatomically shaped contour.

For the quantitative subanalysis of segmentation quality, three methods of ventricular contouring were compared: *(a)* manual, *(b)* fully automated, and *(c)* expert revised. Manual measures were considered the benchmark for comparison with fully automated or expert revised contours. For the comparison between expert revised and automated measures, the expert revised measures were considered the benchmark. This analysis included ED and ES contours for both the LV and RV and was composed of 1283 images. Subanalysis of volumetric correlations, relative volume error, and bias were reported. The Dice coefficient was used to assess contour similarity for LV endocardium, LV epicardium, and RV endocardium for each image slice. An overall Dice coefficient for each ventricle was computed with slices weighted by their contribution to the total ventricular volume. This was performed to normalize relatively small areas, such as the cardiac apex. In addition, regional Dice was computed as an average across multiple slices. Each region (apical, midapical, mid, midbasal, and basal) comprised approximately 20% of the length of the cardiac long axis.

Statistical Analysis

All statistical analysis was performed using a type I error rate (α) of .05 using R 3.4.2 (R Foundation for Statistical Computing, Vienna, Austria) with the packages ggplot2 (29), psych (30), and BlandAltmanLeh (31). Correlations were presented with a Pearson *r* statistic, relative volume error was calculated ([predicted volume – ground truth volume]/ground

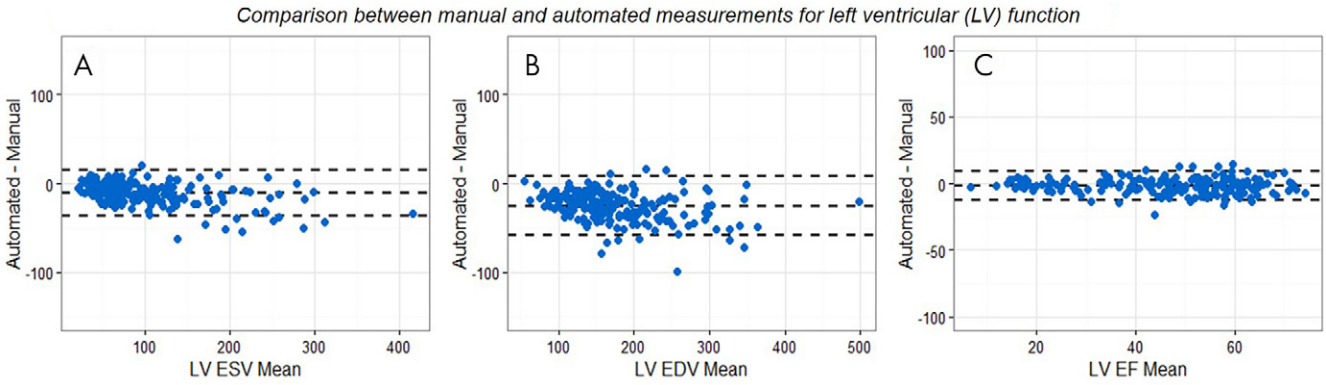

Figure 1: Comparison between manual and automated measurements for left ventricular (LV) function. *A*, Bland-Altman plots compare manual and automated end-systolic volume (ESV), *B*, end-diastolic volume (EDV), and, *C*, ejection fraction (EF) in milliliters and percentages, respectively. Dashed lines indicate average difference and 2 standard deviation limits of agreement. In the LV, the automated metrics were lower for ESV, EDV, and EF.

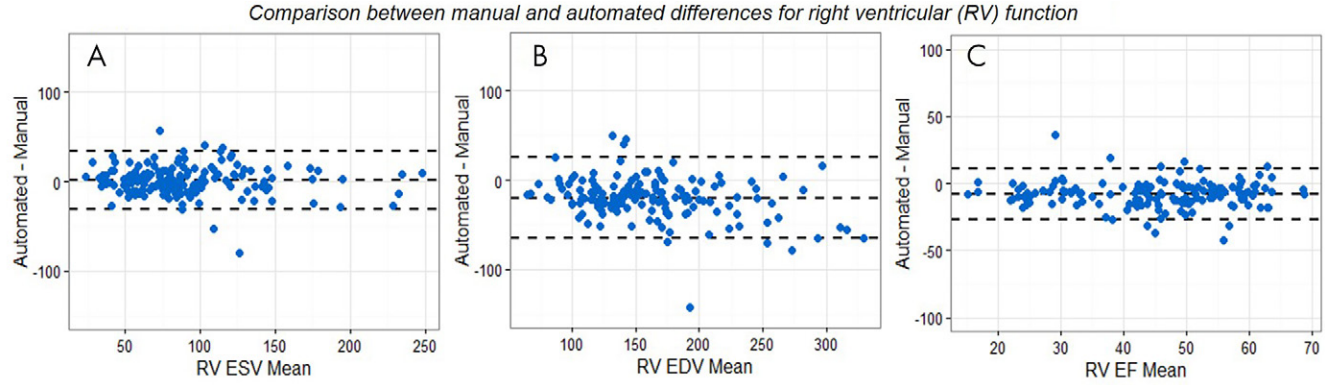

Figure 2: Comparison between manual and automated differences for right ventricular (RV) function. *A*, Bland-Altman plots compare manual and automated end-systolic volume (ESV), *B*, end-diastolic volume (EDV), and *C*, ejection fraction (EF) in milliliters and percentages, respectively. Dashed lines indicate average difference and 2 standard deviation limits of agreement. In the RV, automated measures were, on average, higher than manual for ESV and lower for EDV and EF.

truth volume), Dice coefficients were given as the quotient of similarity between 0 and 1, and Bland-Altman plots indicated average bias and 2 standard deviations as noted.

Results

Contour Generation and Evaluation of Global Measurements

For the primary analysis of performance on all 200 routine cardiac MRI examinations, we observed that the DL-based algorithm generated cardiac contours for most clinical cases. Fully automated contours were generated for 197 (98.5%) for the LV endocardium, 196 (98%) for the LV epicardium, and 160 (80%) for the RV endocardium. Contours that were not generated were not included in the analysis.

Automated LV volumetric measurements showed strong correlation with manual measurements at ESV of $r = 0.99$ ($P < .001$), EDV of $r = 0.97$ ($P < .001$), and EF of $r = 0.94$ ($P < .001$). (Table 2 further details mean differences, relative errors, and 2 standard deviation limits of agreement.) The biases and standard deviations are also shown in Bland-Altman plots in Figure 1. The average difference between the ED mass and ES mass was 1.6 g (mean relative error of 1.5%, and 2 standard deviations at ±30 g).

Automated RV volumetric measurements also showed correlation with manual measurements at ESV of $r = 0.93$ ($P < .001$), EDV of $r = 0.92$ ($P < .001$), and EF of $r = 0.73$ ($P < .001$). (Table 2 further details mean differences, relative errors, and 2 standard deviation limits of agreement.) The biases and standard deviations are shown in Bland-Altman plots in Figure 2.

Subanalysis of the manual, automated, and expert revised contours showed strong correlation (Table 2). Across both ventricles and phases of the cardiac cycle, manual and expert revised measurements had an overall correlation of $r = 0.97$ with a mean difference of 7.1 mL (a mean relative volume error of –4.7%) and 2 standard deviation limits of agreement at ±32.4 mL. DL-derived automated contours agreed with both manual and expert revised contours. When compared with expert revised contours, there was an overall correlation of $r = 0.97$ with mean difference of 7.2 mL (a mean relative volume error of –2.1%) and 2 standard deviation limits of agreement at ±32.6 mL. When compared with manual contours, there was an overall correlation of $r = 0.96$ with mean difference of 14.4 mL (a mean relative volume error of –7.3%) and 2 standard deviation limits of agreement at ±41.4 mL. Bland-Altman plots highlight these comparisons in Figure 3 (LV) and Figure 4 (RV).

Qualitative Comparison of Contours

Automated contours were visually inspected at both ES and ED for errant inclusion or exclusion of apical or basal slices,

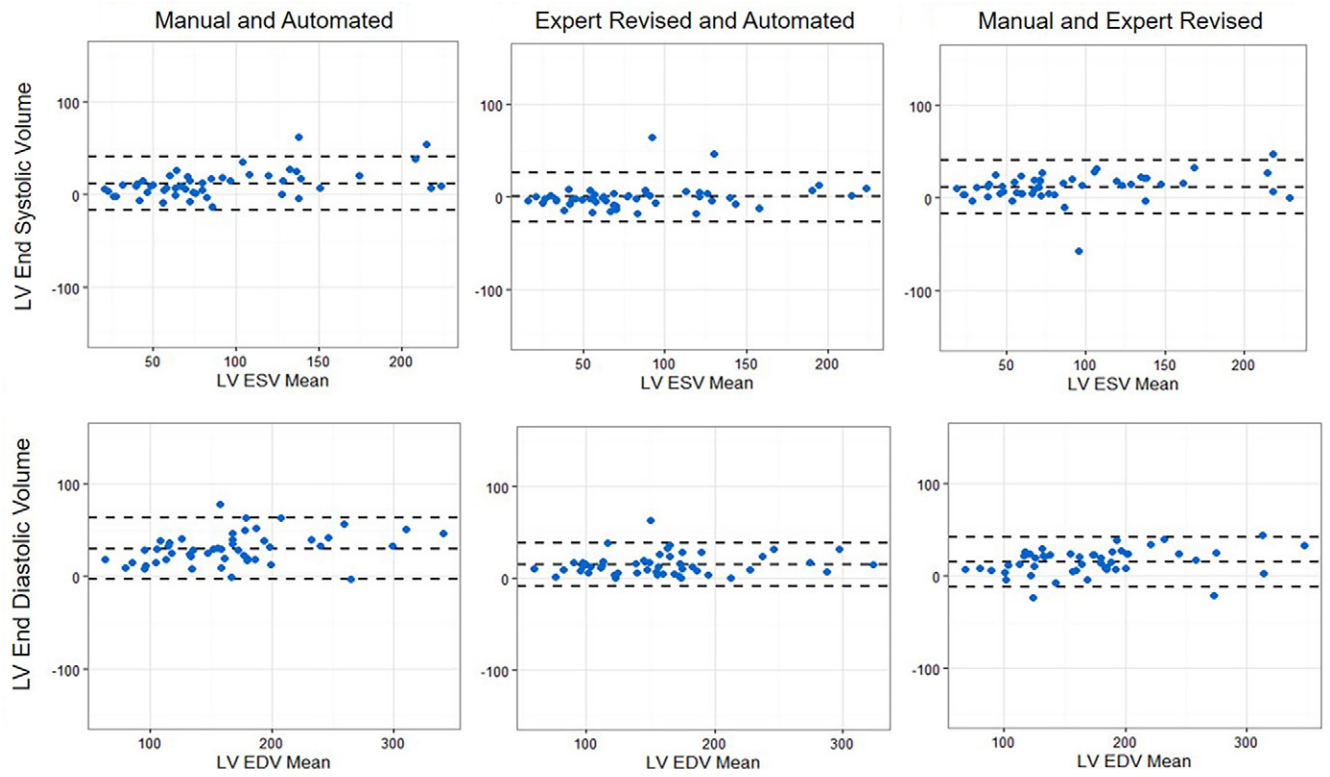

Figure 3: Subanalysis comparing manual, automated, and expert revised measures for left ventricular (LV) end-systolic volume (ESV) and end-diastolic volume (EDV). Dashed lines indicate mean bias and 2 standard deviations range. Overall, manual volumes were larger than automated or expert revised. Expert revised and automated measures were similar at ESV, whereas at EDV the expert revised measures were larger.

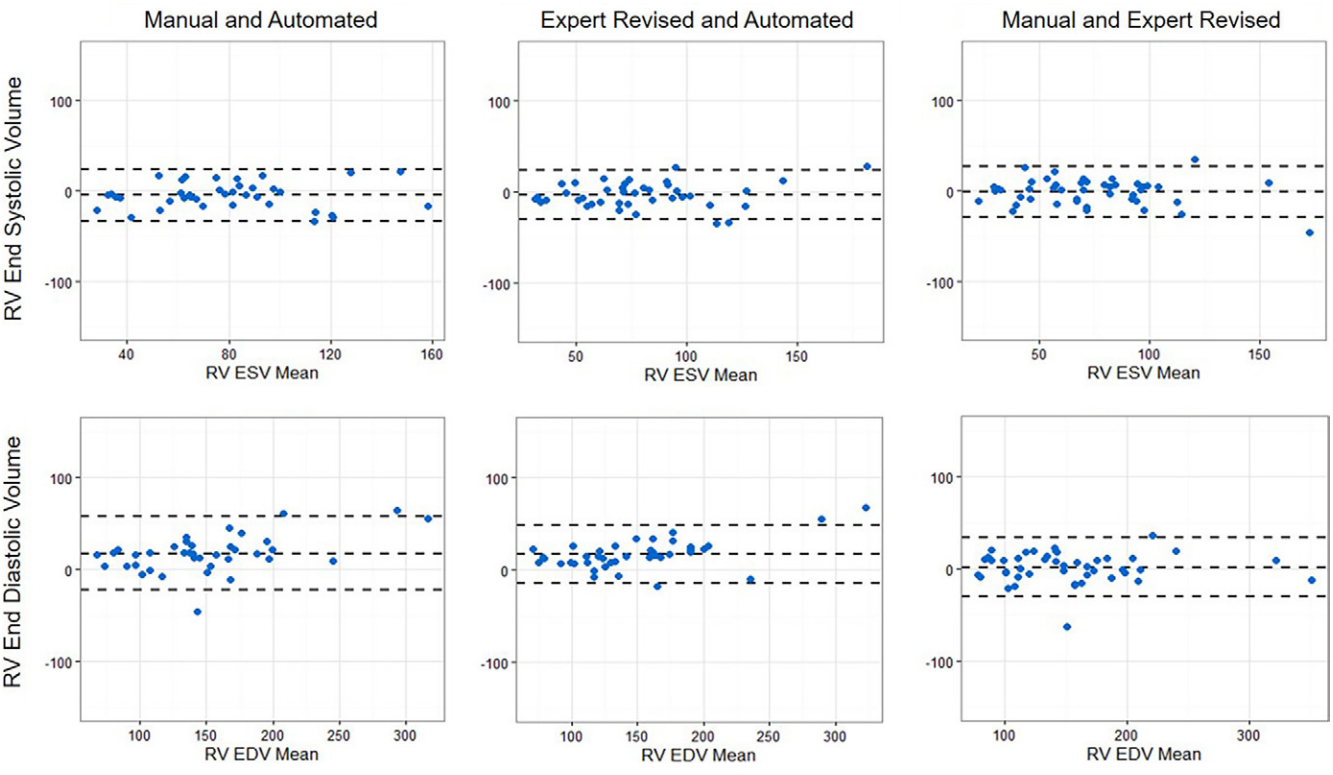

Figure 4: Subanalysis comparing manual, automated, and expert revised measures for right ventricular (RV) end-systolic volume (ESV) and end-diastolic volume (EDV). Dashed lines indicate mean bias and 2 standard deviations range. Manual and expert revised measurements tended to be slightly smaller for the ESV, and manual and expert revised measurements tended to be larger than automated EDVs. Across all groups ESV measurements were the most similar, possibly owing to the inherently larger nature of the EDV and greater variability of basal slice locations at this timepoint.

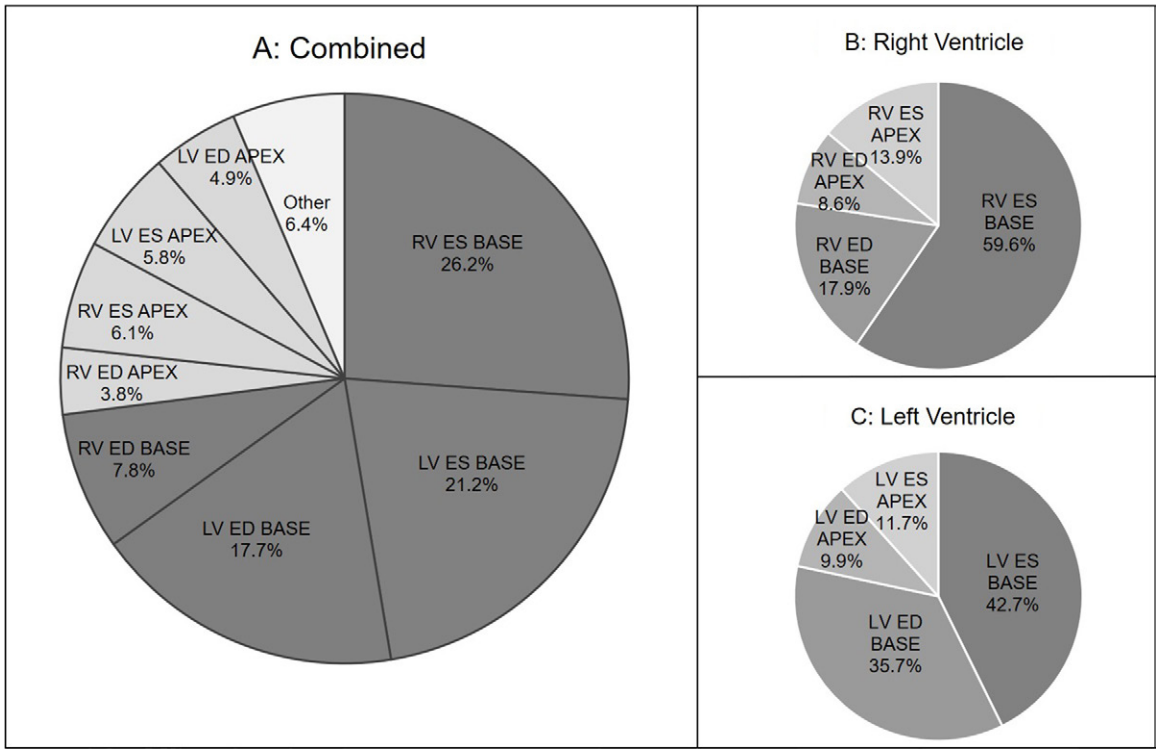

Figure 5: Distribution of automated contour errors. Overall errors for both ventricles are shown in, *A*, and shown by ventricle in, *B*, and *C*. Differences from manual contours were classified as errant inclusion or exclusion of apical or basal slices, or generation of a nonanatomic shape or location of contours. Differences outside the previous criteria were classified as "other." *A*, Chart shows the location (apex or base, and left ventricle [LV] or right ventricle [RV]), and cardiac phase (end systole [ES] or end diastole [ED]) of differences as a percentage of total. The largest proportion was seen at the cardiac base (dark gray slices), an area with variation in contours even among experts. Although the RV ES base had the largest percentage of differences, the LV base (including both endocardium and epicardium) had the largest number of differences overall.

nonanatomic shapes, and nonanatomic locations of contours. Individual classes of errors were counted, allowing for multiple error types simultaneously in a single contour. In 33 of 200 (16.5%) cases, no visually apparent errors were identified. A total of 344 contour errors were classified (Fig 5), and of these, 22 of 344 (6.4%) did not fit one of the classifications and were denoted as "other errors." Errors were distributed evenly between the right side of the heart (151 of 344 [43.9%]) and left side of the heart (171 of 344 [49.7%]). When comparing location, 251 of 344 (73.0%) were at the cardiac basal slices, whereas 71 of 344 (20.6%) were at the cardiac apex. A greater proportion of the errors occurred at ES (204 of 344 [59.3%]) compared with ED (118 of 244 [34.3%]). The individual regions with the greatest proportion of errors were the RV base at ES (90 of 344 [26.2% of total]) and the LV base at ES (73 of 244 [21.2% of total]). Examples of errant contours are shown in Figure E1 (supplement).

Quantitative Comparison of Contours

The Dice metric compares similarity between two contour areas with a range between 0 (indicating no overlap) and 1 (indicating perfect overlap). When comparing the manual with expert revised segmentations, the weighted average Dice coefficient for all contours at both phases of the cardiac cycle was 0.94. Comparison between manual and automated segmentations had an overall weighted average Dice coefficient of 0.92, and expert revised segmentations compared with automated segmentations had an overall weighted average Dice coefficient of 0.95. When comparing the interaction between contour and region for the manual, automated, and expert revised Dice metrics, a two-way analysis of variance was not significant for LV endocardium, LV epicardium, or RV endocardium at ED ($P = .23$, $P = .95$, and $P = .61$, respectively), or ES ($P = .76$, $P = .71$, $P = .92$, respectively), indicating that the Dice is not significantly different between these groups. The average Dice between these groups for each cardiac region and phase of the cardiac cycle is shown in Table 3 and Figure 6, with examples of manual, expert revised, and automated contours shown in Figure E2 (supplement). In aggregate, Dice metrics were similar across systole and diastole. When examined by contour, the LV epicardium had the most consistently high Dice values, and the RV endocardium showed the most variability in contours with larger Dice variation and lower average scores across group comparisons. Consistently high Dice values were calculated in the midventricular segments, indicating highly congruent contours in these regions among all groups. The region with the greatest variation across both systole and diastole was the cardiac apex. Of note, while the apical region was defined as 20% of cardiac length on the short-axis images, it accounted for an average of 8.2% of overall cardiac volume in diastole and 3.1% of overall cardiac volume in systole. Therefore, while the apex had the largest amount of contour variability, the impact on overall volume calculation was relatively small.

Table 3: Dice Correlations across Groups

Contour	Phase	Average Dice	Apical	Mid-Apical	Mid	Mid-Basal	Basal
			Manual and Expert Revised Segmentations				
LV endocardium	ESV	0.91	0.79	0.87	0.92	0.92	0.89
	EDV	0.91	0.86	0.92	0.93	0.92	0.88
LV epicardium	ESV	0.95	0.93	0.95	0.96	0.96	0.95
	EDV	0.95	0.94	0.96	0.96	0.95	0.93
RV endocardium	ESV	0.86	0.77	0.81	0.87	0.9	0.8
	EDV	0.88	0.79	0.86	0.9	0.9	0.84
			Manual and Automated Segmentations				
LV endocardium	ESV	0.91	0.81	0.87	0.92	0.92	0.88
	EDV	0.92	0.88	0.93	0.94	0.94	0.88
LV epicardium	ESV	0.95	0.91	0.94	0.96	0.95	0.94
	EDV	0.94	0.92	0.95	0.95	0.95	0.9
RV endocardium	ESV	0.85	0.81	0.7	0.84	0.9	0.76
	EDV	0.88	0.74	0.82	0.87	0.91	0.85
			Expert Revised and Automated Segmentations				
LV endocardium	ESV	0.95	0.85	0.94	0.96	0.95	0.92
	EDV	0.95	0.92	0.97	0.97	0.96	0.89
LV epicardium	ESV	0.95	0.91	0.96	0.96	0.95	0.95
	EDV	0.96	0.95	0.97	0.96	0.97	0.92
RV endocardium	ESV	0.9	0.9	0.77	0.88	0.92	0.83
	EDV	0.92	0.86	0.91	0.92	0.94	0.89

Note.—The overall weighted averages for manual and expert revised segmentations, manual and automated segmentations, and expert revised and automated segmentations were 0.94, 0.92, and 0.95, respectively. EDV = end-diastolic volume, ESV = end-systolic volume, LV = left ventricle, RV = right ventricle.

Discussion

In this study, we showed that a DL algorithm can perform both RV and LV volumetry on cardiac MRI data from most patients in clinical practice. There was strong correlation between volumetric measurements obtained by fully manual and fully automated approaches. Remarkably, the differences between manual and automated cardiac volumetry were within the range seen between expert readers (5). Suinesiaputra et al previously compared cardiac MRI measurements from readers at multiple institutions from MRI examinations with a variety of pathologies and image quality and noted an error range from −32.9 to 41.2 mL relative to the consensus for LV ESV (5). In our study, we observed that between manual and automated contours, limits of agreement for LV ESV fell within this range. These results suggest that incorporation of fully automated contours into the diagnostic workflow may help to reduce the variability typically seen between readers.

When we explored the overlap of individual segmentations on a regional basis, the DL algorithm performed better in the midventricular slices than at the base and apex. The basal and apical segments have also been shown to have the most variability when segmentations were attempted by other algorithms and generated the greatest variability between expert readers (21,32). For example, multiple prior studies have reported high interobserver contour variation at the cardiac base (5,33–35). Bonnemains et al examined reader variability through the RV short axis and found the basal RV accounted for 70%–80% of total variability among expert readers (36). The authors proposed that the imaging plane may make it difficult to agree on the extent of the RV because the position of the valves adds uncertainty to the volume (36). The difficulty that expert readers and DL algorithms have with this area may therefore be, in part, the result of ambiguity of right side of the heart boundaries on the basal images themselves.

The DL algorithm tested in our study performed well compared with other recent machine learning–based RV segmentation methods. For example, an algorithm proposed by Avendi et al showed similarly strong correlation with ground truth measurements (0.99 for end systolic and 0.98 for end diastolic) and had a Dice area similarity coefficient of 0.82 (9). Although the RV volumes in that study had a stronger correlation, Dice coefficients were lower. This finding may be related to the scope and variability of the image dataset and quality of the ground truth segmentations near the tricuspid valve (9,23,24). Further studies are needed to investigate the performance of different candidate algorithms on common clinical datasets, such as those that organizations like the American College of Radiology are developing in shared repositories for imaging data.

There were several limitations to the current study. It is possible that a wider range of ground truth segmentations from multiple readers could alter the performance metrics, and it is clear from prior studies that there can be disagreement on proper segmentation of the basal and apical slices between expert readers. The ground truth segmentations for this study were

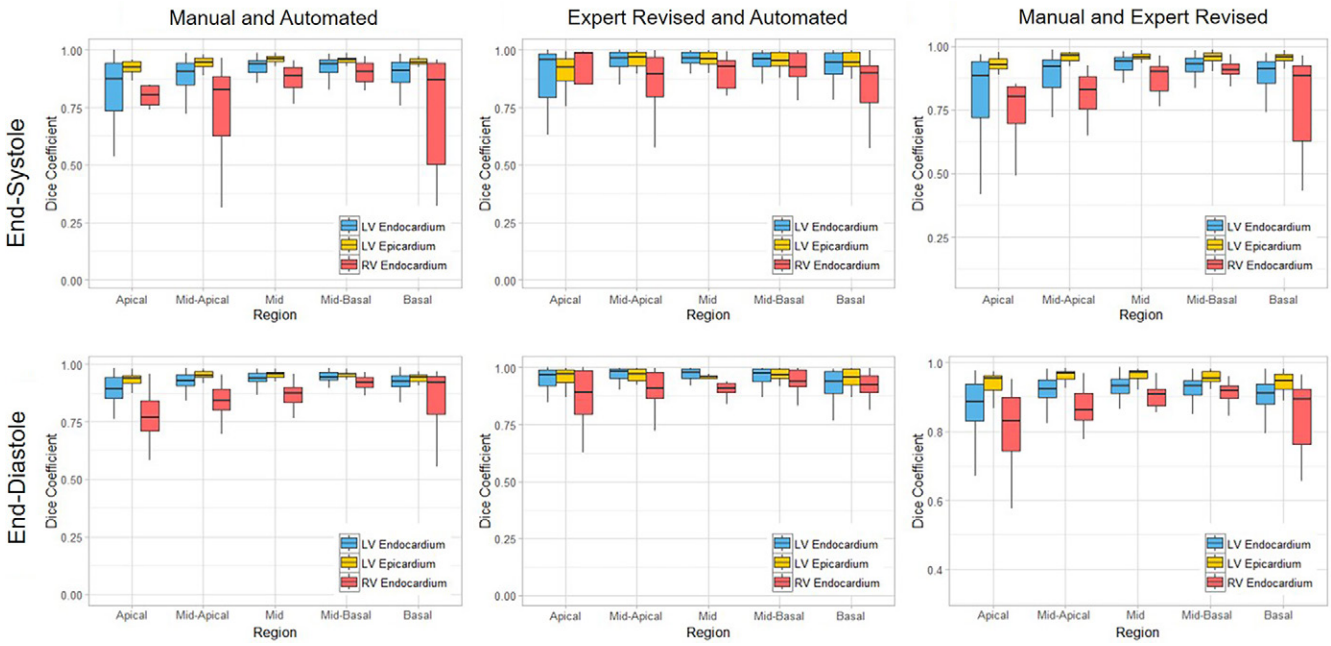

Figure 6: Dice coefficients by cardiac region (apex to base). Dice coefficients, a measurement of area similarity, were calculated to compare automated with manual and expert revised cardiac contours. Dice coefficients are shown for end diastole and systole. Data are presented for the left ventricle (LV) endocardial, LV epicardial, and right ventricle (RV) endocardial contours. Dice coefficients are most consistent and highest for the LV epicardium, with greatest variation for the RV endocardium. Dice scores were consistent and highest through the midventricular slices, with larger variation at the apical and basal slices.

taken from measurements performed in the context of clinical practice. We acknowledge that variance may occur between the different clinical readers; however, comparison of algorithm performance with a clinical endpoint was chosen to reflect the expected use scenario. Second, the initial version of the DL algorithm displayed a contour based on an internal confidence threshold, which prevented it from generating contours for all cases. This may not necessarily represent a fundamental limitation of convolutional neural networks, but rather a design decision for clinical usability. Because algorithm contours below this threshold were not available for study inclusion, the algorithm performance on RV segmentation of all cases is likely lower on average than found here. Future work could be directed toward optimizing this confidence threshold. Third, the indications, populations, and equipment used for patients referred for cardiac MRI may vary between institutions. In particular, the performance of the algorithm studied here may not be as robust with cases of complex congenital heart disease, or real-time cases which were not present in the population analyzed. Future directions could include testing on patients with congenital heart defects and gathering additional studies for subanalyses of patients with specific pathologies, in addition to examining the performance on real-time cases. In addition, we recognize the limitation of using cases from a single MRI scanner. This study represents an independent test of the algorithm because none of its training data came from our institution. However, to further assess algorithm performance and generalizability, future testing should be performed using equipment from multiple MRI vendors and multiple sequences.

Cardiac MRI is a model system for the study of clinical implementation of DL algorithms, particularly because quantitative volumetric measurements are a routine part of clinical practice. DL algorithms have the potential to increase accessibility of cardiac MRI by decreasing the time necessary to obtain volumetric measurements. Implementation of DL algorithms in practice may ultimately benefit clinical practice by allowing physicians to focus on higher order tasks in patient management rather than manual cardiac contouring. In this study, we showed that a DL–based algorithm can segment the RV and LV in a manner similar to that of expert readers. We believe that the combination of DL automation and specialist oversight can enhance patient care by streamlining quantitative interpretation, particularly as these algorithms continue to improve.

Acknowledgments: The authors would like to thank Kevin Blansit, Kang Wang, and Naeim Bahrami for their kind support and insightful discussions.

Author contributions: Guarantors of integrity of entire study, T.A.R., A.H.; 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.A.R., A.H.; clinical studies, T.A.R., A.H.; statistical analysis, T.A.R., E.M.M.; and manuscript editing, T.A.R., A.H.

Disclosures of Conflicts of Interest: T.A.R. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives RSNA Resident Research grant; institution (UC San Diego) receives money for patent for MRI plane prescription software; author part of research residency at UCSD funded by NIH T32 training grant EB005970. Other relationships: disclosed no relevant relationships. **E.M.M.** Activities related to the present article: institution receives grant from National Institutes of Health (T32 HL 105373). Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **D.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Arterys. Other relationships: disclosed no relevant relationships. **A.H.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is consultant/cofounder for Arterys; institution receives grant from GE Healthcare and Bayer; author paid for lectures by GE Healthcare and Bayer; institution has patent (Stan-

ford University, UC San Diego, and Arterys); author receives royalties from Stanford University; author is founding shareholder in Arterys; author receives travel accommodations from GE Healthcare and Arterys. Other relationships: disclosed no relevant relationships.

References

1. Childs H, Ma L, Ma M, et al. Comparison of long and short axis quantification of left ventricular volume parameters by cardiovascular magnetic resonance, with ex-vivo validation. J Cardiovasc Magn Reson 2011;13(1):40.
2. Hendel RC, Patel MR, Kramer CM, et al. ACCF/ACR/SCCT/SCMR/ASNC/NASCI/SCAI/SIR 2006 appropriateness criteria for cardiac computed tomography and cardiac magnetic resonance imaging: a report of the American College of Cardiology Foundation Quality Strategic Directions Committee Appropriateness Criteria Working Group, American College of Radiology, Society of Cardiovascular Computed Tomography, Society for Cardiovascular Magnetic Resonance, American Society of Nuclear Cardiology, North American Society for Cardiac Imaging, Society for Cardiovascular Angiography and Interventions, and Society of Interventional Radiology. J Am Coll Cardiol 2006;48(7):1475–1497.
3. Pesenti-Rossi D, Peyrou J, Baron N, et al. Cardiac MRI: technology, clinical applications, and future directions [in French]. Ann Cardiol Angeiol (Paris) 2013;62(5):326–341.
4. Flett AS, Westwood MA, Davies LC, Mathur A, Moon JC. The prognostic implications of cardiovascular magnetic resonance. Circ Cardiovasc Imaging 2009;2(3):243–250.
5. Suinesiaputra A, Bluemke DA, Cowan BR, et al. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. J Cardiovasc Magn Reson 2015;17(1):63.
6. Schulz-Menger J, Bluemke DA, Bremerich J, et al. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) board of trustees task force on standardized post processing. J Cardiovasc Magn Reson 2013;15(1):35.
7. Caudron J, Fares J, Vivier PH, Lefebvre V, Petitjean C, Dacher JN. Diagnostic accuracy and variability of three semi-quantitative methods for assessing right ventricular systolic function from cardiac MRI in patients with acquired heart disease. Eur Radiol 2011;21(10):2111–2120.
8. Suinesiaputra A, Cowan BR, Al-Agamy AO, et al. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. Med Image Anal 2014;18(1):50–62.
9. Avendi MR, Kheradvar A, Jafarkhani H. Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach. Magn Reson Med 2017;78(6):2439–2448.
10. Beerbaum P, Barth P, Kropf S, et al. Cardiac function by MRI in congenital heart disease: impact of consensus training on interinstitutional variance. J Magn Reson Imaging 2009;30(5):956–966.
11. Cocosco CA, Niessen WJ, Netsch T, et al. Automatic image-driven segmentation of the ventricles in cardiac cine MRI. J Magn Reson Imaging 2008;28(2):366–374.
12. Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. Med Image Anal 2011;15(2):169–184.
13. Peng P, Lekadir K, Gooya A, Shao L, Petersen SE, Frangi AF. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. MAGMA 2016;29(2):155–195.
14. Slomka PJ, Dey D, Sitek A, Motwani M, Berman DS, Germano G. Cardiac imaging: working towards fully-automated machine analysis & interpretation. Expert Rev Med Devices 2017;14(3):197–212.
15. Pednekar AS, Muthupillai R, Cheong B, Flamm SD. Automatic computation of left ventricular ejection fraction from spatiotemporal information in cine-SSFP cardiac MR images. J Magn Reson Imaging 2008;28(1):39–50.
16. Zhuang X, Hawkes DJ, Crum WR, et al. Robust registration between cardiac MRI images and atlas for segmentation propagation. In: Reinhardt JM, Pluim JPW, eds. Proceedings of SPIE: Medical Imaging 2008—Image Processing. Vol 6914. Bellingham, Wash: International Society for Optics and Photonics, 2008; 691408.
17. Lorenzo-Valdés M, Sanchez-Ortiz GI, Elkington AG, Mohiaddin RH, Rueckert D. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. Med Image Anal 2004;8(3):255–265.
18. Bahrami N, Retson T, Blansit K, Wang K, Hsiao A. Automated selection of myocardial inversion time with a convolutional neural network: spatial temporal ensemble myocardium inversion network (STEMI-NET). Magn Reson Med 2019;81(5):3283–3291.
19. Lieman-Sifry J, Le M, Lau F, Sall S, Golden D. FastVentricle: Cardiac Segmentation with ENet. ArXiv [preprint] https://arxiv.org/abs/1704.04296. Posted April 13, 2017. Accessed November 2017.
20. Narayan T. Automated Left Ventricle Segmentation in Cardiac MRIs using Convolutional Neural Networks. http://cs231n.stanford.edu/reports/2016/pdfs/317_Report.pdf. Accessed September 2017.
21. Tao Q, Yan W, Wang Y, et al. Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 2019;290(1):81–88.
22. Tan LK, McLaughlin RA, Lim E, Abdul Aziz YF, Liew YM. Fully automated segmentation of the left ventricle in cine cardiac MRI using neural network regression. J Magn Reson Imaging 2018;48(1):140–152.
23. Petitjean C, Zuluaga MA, Bai W, et al. Right Ventricle Segmentation from Cardiac MRI: A Collation Study. http://www.litislab.fr/wp-content/uploads/2014/07/RSCVMedIApaper_finalversion.pdf. Published 2014. Accessed May 2018.
24. Ringenberg J, Deo M, Devabhaktuni V, Berenfeld O, Boyers P, Gold J. Fast, accurate, and fully automatic segmentation of the right ventricle in short-axis cardiac MRI. Comput Med Imaging Graph 2014;38(3):190–201.
25. Suinesiaputra A, Cowan BR, Finn JP, et al. Left ventricular segmentation challenge from cardiac MRI: a collation study. In: Camara O, Konukoglu E, Pop M, Rhode K, Sermesant M, Young A, eds. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. STACOM 2011. Lecture Notes in Computer Science, vol 7085. Berlin, Germany: Springer, 2012; 88–97.
26. Radau P. Cardiac Atlas Project - Sunnybrook Cardiac Data. http://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/. Accessed May 2018.
27. Radau P, Lu Y, Connelly K, Paul G, Dick AJWG. Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI. The MIDAS Journal – Cardiac MR Left Ventricle Segmentation Challenge. http://hdl.handle.net/10380/3070. Accessed May 2018.
28. Petersen SE, Matthews PM, Bamberg F, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. J Cardiovasc Magn Reson 2013;15(1):46.
29. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag, 2009.
30. Revelle W. psych: Procedures for Personality and Psychological Research. Evanston, Ill: Northwestern University. https://cran.r-project.org/package=psych. Published 2017. Last update December 2019.
31. Lehnert B. BlandAltmanLeh. https://cran.r-project.org/web/packages/BlandAltmanLeh/index.html. Published 2015. Accessed September 2017.
32. 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.
33. Karamitsos TD, Hudsmith LE, Selvanayagam JB, Neubauer S, Francis JM. Operator induced variability in left ventricular measurements with cardiovascular magnetic resonance is improved after training. J Cardiovasc Magn Reson 2007;9(5):777–783.
34. Groth M, Muellerleile K, Klink T, et al. Improved agreement between experienced and inexperienced observers using a standardized evaluation protocol for cardiac volumetry and infarct size measurement. Rofo 2012;184(12):1131–1137.
35. Catalano O, Antonaci S, Opasich C, et al. Intra-observer and interobserver reproducibility of right ventricle volumes, function and mass by cardiac magnetic resonance. J Cardiovasc Med (Hagerstown) 2007;8(10):807–814.
36. Bonnemains L, Mandry D, Marie PY, Micard E, Chen B, Vuissoz PA. Assessment of right ventricle volumes and function by cardiac MRI: quantification of the regional and global interobserver variability. Magn Reson Med 2012;67(6):1740–1746.