Circulation: Cardiovascular Imaging

ORIGINAL ARTICLE

A Multicenter, Scan-Rescan, Human and Machine Learning CMR Study to Test Generalizability and Precision in Imaging Biomarker Analysis

See Editorial by Colletti

Anish N. Bhuva, MBBS et al

BACKGROUND: Automated analysis of cardiac structure and function using machine learning (ML) has great potential, but is currently hindered by poor generalizability. Comparison is traditionally against clinicians as a reference, ignoring inherent human inter- and intraobserver error, and ensuring that ML cannot demonstrate superiority. Measuring precision (scan:rescan reproducibility) addresses this. We compared precision of ML and humans using a multicenter, multi-disease, scan:rescan cardiovascular magnetic resonance data set.

METHODS: One hundred ten patients (5 disease categories, 5 institutions, 2 scanner manufacturers, and 2 field strengths) underwent scan:rescan cardiovascular magnetic resonance (96% within one week). After identification of the most precise human technique, left ventricular chamber volumes, mass, and ejection fraction were measured by an expert, a trained junior clinician, and a fully automated convolutional neural network trained on 599 independent multicenter disease cases. Scan:rescan coefficient of variation and 1000 bootstrapped 95% CIs were calculated and compared using mixed linear effects models.

RESULTS: Clinicians can be confident in detecting a 9% change in left ventricular ejection fraction, with greater than half of coefficient of variation attributable to intraobserver variation. Expert, trained junior, and automated scan:rescan precision were similar (for left ventricular ejection fraction, coefficient of variation 6.1 [5.2%–7.1%], *P*=0.2581; 8.3 [5.6%–10.3%], *P*=0.3653; 8.8 [6.1%–11.1%], *P*=0.8620). Automated analysis was 186× faster than humans (0.07 versus 13 minutes).

CONCLUSIONS: Automated ML analysis is faster with similar precision to the most precise human techniques, even when challenged with real-world scan:rescan data. Assessment of multicenter, multi-vendor, multi-field strength scan:rescan data (available at www.thevolumesresource.com) permits a generalizable assessment of ML precision and may facilitate direct translation of ML to clinical practice.

The full author list is available on page 10.

Key Words: artificial intelligence ■ image processing ■ left ventricular remodeling ■ magnetic resonance imaging, cine ■ ventricular function

© 2019 The Authors. *Circulation: Cardiovascular Imaging* is published on behalf of the American Heart Association, Inc., by Wolters Kluwer Health, Inc. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial-NoDerivs License, which permits use, distribution, and reproduction in any medium, provided that the original work is properly cited, the use is noncommercial, and no modifications or adaptations are made.

https://www.ahajournals.org/journal/circimaging

Downloaded from http://ahajournals.org by on September 24, 2019

CLINICAL PERSPECTIVE

Left ventricular ejection fraction and mass remain key imaging biomarkers and are used daily for clinical decision-making and as clinical trial outcome measures. While cardiovascular magnetic resonance imaging to measure left ventricular ejection fraction is performed at high resolution, clinician analysis is remarkably variable. Automated techniques using machine learning may offer time-saving and improved confidence in absolute values, but they should be demonstrably generalizable and precise (repeatable) before widespread adoption. To address this, a multicenter, multi-vendor, multi-field strength, multi-disease cardiovascular magnetic resonance resource of 110 patients undergoing repeat imaging in a short time-frame was assembled. This was analyzed by an expert, a trained junior clinician (using five different techniques), and a fully automated convolutional neural network. This showed that clinicians can be confident in detecting a 9% change in left ventricular ejection fraction or a 20 g change in LV mass. This will be difficult to improve for clinicians because the greatest source of human error was attributable to the observer rather than modifiable factors. Having understood these errors, a convolutional neural network was trained on separate multicentre data for automated analysis and was successfully generalizable to the real-world cardiovascular magnetic resonance data. Precision was similar to human analysis, and performance was 186× faster. Automated cardiovascular magnetic resonance analysis should, therefore, be adopted globally to gain from time-saving and standardization benefits. The real-world benchmarking resource has been made available, detailed at www.thevolumesresource.com.

Left ventricular ejection fraction (LVEF) and mass (LVM) remain key imaging biomarkers and are used daily for clinical decision-making and as clinical trial outcome measures.[1,2] Absolute values guide pharmacotherapy, device therapy, and surgical intervention. Although it is important that measurement is accurate against some putative reference- or at least that any bias is known, it is measurement precision (repeatability) that determines the clinical smallest detectable difference with time or treatment and the sample size of clinical trials.[3] Cardiovascular magnetic resonance (CMR) imaging is the reference standard imaging modality to assess LV structure and function, and image acquisition is largely standardized through international consensus guidelines.[4] In contrast, there is less agreement about analysis techniques where significant variation exists between inclusion/exclusion of papillary muscles, trabeculae, and use of edge detection methods despite the CMR community desire for consistency and precision.

LVEF and LVM measurement variation arise from many sources including on-target changes with disease or intervention, off-target unavoidable biological variation (eg, heart rate and volume status), and avoidable intraobserver, inter-observer, inter-study, inter-center variation. These include noise (eg, intraobserver variation) and bias (one observer may systematically detect an edge differently to another),[5,6] but relative contributions of each error source are not known.

Training programs and semi-automated contouring speed up segmentation and improve inter-observer agreement, but techniques vary considerably.[7–9] Automated analysis via machine learning (ML) approaches using deep learning neural networks show potential,[10,11] and could remove this intra- and inter-observer variation. Currently, ML algorithms are tested by direct comparison with human expert observers as the reference standard, however this ignores sources of human error and means that ML techniques are unable to demonstrate superiority over human techniques. Precision can only be assessed using a test:retest data set- this requires a significant sized patient cohort to be scanned twice in an identical fashion within an interval short enough to effectively exclude variation in disease biology.[5,12,13] For generalizability, this should be done across multiple sites and platforms.

A multi-scanner, multicenter, health and disease precision (scan-rescan) CMR data set resource for use as a tool to measure human and ML LVEF, and LVM analysis performance was collated. This resource was then used to quantify CMR precision and different sources of human error (scan acquisition, observer experience, level of automation) using multiple analysis techniques. Having understood error sources from human approaches, a deep learning convolutional neural network was trained on a large multiscanner multicenter disease cohort and explored human and ML performance. It was hypothesized that greater clinician experience and semi-automated contouring would improve human precision, and that an automated technique would have superior performance overall.

METHODS

Data Availability

The authors declare that all supporting data are available within the article and its in the Data Supplement. Details of scan-rescan data set availability are at www.thevolumesresource.com (Validation Of Left ventricular Myocardial and Endocardial Segmentation resource), intended for those wishing to benchmark future automated analysis approaches.

Study Population

The scan-rescan CMR parameters for precision assessment are outlined in Table I in the Data Supplement. In brief, paired scans were obtained from 5 United Kingdom institutions (Barts

Downloaded from http://ahajournals.org by on September 24, 2019

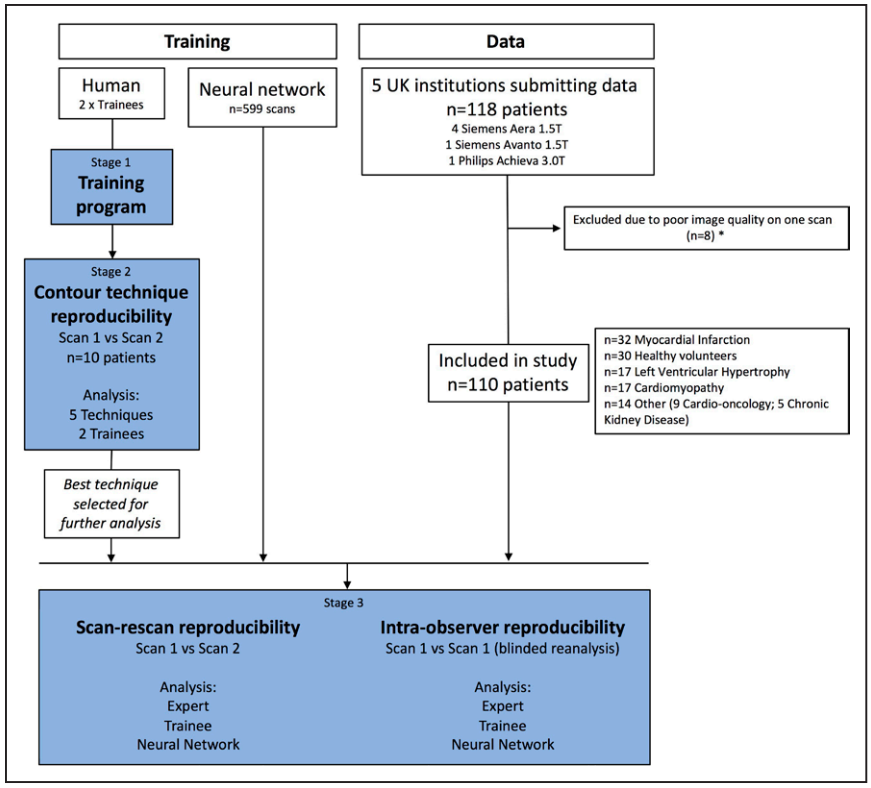

Figure 1. Flow diagram illustrating study recruitment and analysis.
Shaded blue boxes represent study outcome measures. *Missing slices or significant breathing artifact on one scan.

Heart Centre, University Hospitals Bristol, Leeds Teaching Hospitals, University College London Hospital, and University Hospitals Birmingham NHS Trusts) with 6 different MRI scanners of 2 field strengths (1.5T, 3T), 2 manufacturers (Siemens, Philips), and 3 models (Aera, Achieva, Avanto) representing the clinical spectrum (health, dilatation, hypertrophy, regional disease, n=118). Each institution obtained local approval via the United Kingdom National Research Ethics Service; the study conformed to the principles of the Helsinki Declaration, and all subjects gave written informed consent. Inclusion criteria were patients over age 18 years undergoing CMR with balanced steady-state free precession cine imaging on 2 occasions within a time-frame where biological change was not anticipated. Scans:rescans were acquired either both before or after gadolinium based contrast administration, using the same protocol, as per the international guidelines on scan acquisition.[4] Scans acquired on the same day involved removing the patient from the table and performing repeat isocenter positioning. Exclusion criteria included patients with a cardiac implantable electronic device, significant arrhythmia (atrial fibrillation or ectopy) during the scan, claustrophobia or inability to breath-hold. Allometric data were collected before the scan, and body surface area (BSA) was calculated using the Mosteller formula, $\left(\frac{\text{Height x Weight}}{3600}\right)^{1/2}$. Diagnoses were provided by the recruiting center.

CMR Scan-Rescan and Intraobserver Reproducibility

Each data set consisted of cine imaging in at least 2 long-axis orientations and a complete short-axis stack. Both scans per patient were assigned separate, randomly generated, 4-digit identification codes for blinded scan and rescan assessment. The first scan was also duplicated and assigned a separate identification code for assessment of blinded intraobserver reassessment. Data sets were excluded (n=8) if there were missing slices or unacceptable quality on one or other acquisition judged by an expert observer (Dr Moon).

Clinician Analysis

Images were analyzed by an expert (Dr Moon) with greater than 15 years experience and 2 cardiology trainees (Drs Ye and Lau) with less than one year of experience reporting CMR. With 5 human analysis techniques, variation in performance was expected. A 3 stage process was therefore designed, Figure 1.

Stage One

Two trainees undertook a training/standardization program over one month. Both were initially Society of Cardiovascular Magnetic Resonance level 1 accredited, and they had contoured ≈100 and 700 scans with senior clinicians respectively for the 2 observers. Contouring feedback was provided by 2 experts (Drs Moon and Manisty) and standardized instructions created (consensus—based on local practice and informed by known international standard operating procedures within UK Biobank and MESA),[14]—see tutorial video and standard operating procedures in Methods in the Data Supplement. Fifteen (different) studies ranging in difficulty were contoured and then recontoured a month later to assess training impact.

Stage Two

The 2 trainees each analyzed 10 scan-rescans of patients representing different pathologies using 5 techniques (total 200 complete LVs contoured; Methods in the Data Supplement and Figure 2). Techniques were (1) free-hand fully manual contouring; (2) visual thresholding of the blood-myocardial

Downloaded from http://ahajournals.org by on September 24, 2019

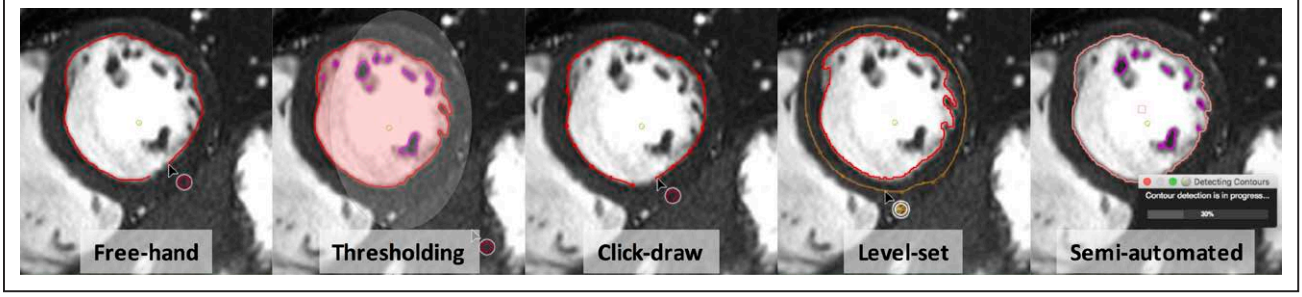

Figure 2. Manual and semi-automated techniques used to segment the endocardial border.
From left to right segmentation techniques become increasingly semi-automated: fully manual; visual signal intensity-based thresholding; manual point series; mid-myocardial contour to initialize level-set segmentation; and blood-pool centering to initialize semi-automated segmentation.

border; (3) curve fitting between points on the endocardium (click-draw); (4) level-set segmentation initialized by a mid-myocardial contour; and (5) semi-automated contouring after manual centering of the LV blood-pool. To minimize the impact of learning, at least 24 hours were left between analysis using the separate techniques, and scans were analyzed in a random order for each technique.

Stage Three

This transitioned the most precise technique from stage 2 for both the expert and trainees to analyze the entire data set of 110 patients (scan A, scan B, blinded scan A again). This totaled 330 LV volumes by expert; 330 by trainees—by this time the trainees were indistinguishable in performance so they acted as one observer, dividing work; and 330 by the automated neural network). Analysis time was measured for a sample of 50 scans for one trainee and for all automated analyses.

All analyses took place using a bespoke prototype of *CVI42* (Release 5.3.8 [720], Circle, Calgary, Canada) to include a novel level-set segmentation technique previously developed and implemented for fractal analysis and modified in this study for endocardial contouring.[15] End diastolic (ED) and end systolic (ES) phases were defined as the largest and smallest long-axis ventricular volumes visually. Contiguous short-axis slices were delineated in ED (endocardium first then epicardium) and ES (endocardium) to derive LV end diastolic volume (EDV), end systolic volume (ESV), stroke volume (SV), ejection fraction (LVEF) and LVM, with allometric scaling using body surface area. To address basal slice variability, blood volume was included if there was over 50% of LV myocardium surrounding blood-pool, and a long-axis atrioventricular plane correction was used. The left ventricular outflow tract was included in the blood volume.

Automated Neural Network Training and Analysis

An automated 2-dimensional deep fully convolutional neural network was previously developed to predict LV endocardial and epicardial contours at ED and ES from an input CMR steady-state free precession short-axis stack.[10] The network was previously trained on 4875 subjects from the UK Biobank, however performance was not easily generalizable to multicenter, multi-disease data. So, the network was trained from scratch on 599 multicenter, multi-scanner data sets of patients with severe aortic stenosis, as described elsewhere.[16] Although this cohort represents one primary disease, extensive comorbidity (hypertension ≈53%, diabetes mellitus ≈22%, coronary artery disease ≈29%), and the ventricular response (≈50% with focal scar, ≈60% with hypertrophy [3 different subtypes], ≈20% with impairment) made it representative of human cardiac disease in general. These 599 scans (comprising ≈13 cines, each 25 frames 195 000 images) were annotated at ED and ES by an expert observer (Dr Moon). Annotations were performed after the standardized post-processing guidelines above. Papillary muscles and trabeculations were included in the LV blood-pool. Training of the network took 8 hours 40 minutes on a Nvidia Titan X GPU.

Statistics

Data were analyzed in R (R foundation, Vienna, Austria) using RStudio Server version 0.98 (Boston, Mass). All continuous variables are expressed as mean±SD or median (interquartile range) for skewed data. Categorical variables are expressed as frequencies in percent. Multiple groups were compared using one-way ANOVA.

For inter-observer agreement, an intraclass correlation coefficient (Lin's concordance correlation coefficient) was used which reflects both agreement and deviation from the line of perfect concordance; <0.2=poor, >0.8=excellent agreement.[17]

To quantify reproducibility, 3 metrics were used: the absolute difference between scans, Bland-Altman limits of agreement, and within-subject Coefficient of variation (CV).[18] The within-subject variance was bootstrapped (1000 bootstraps) to estimate a 95% CI of the CV.

For a head-to-head comparison of scan-rescan precision between techniques or operators, linear mixed effects regression models were used, which account for multiple observers analyzing multiple measures per subject. Models were run separately for each LV metric (EDV, ESV, SV, LVEF, or LVM) as the dependent variable. To assess training, the dependent variable was the difference between expert and trainee, and the fixed effect was training category (before/after). To assess technique, the dependent variable was the LV metric, and the fixed effects were technique and scan-rescan category (1 or 2). An interaction between technique and scan category was used to assess scan-rescan precision for each technique.[19] To assess different operators, the dependent variable was the LV metric, and the fixed effects were operator, scan-rescan category and the interaction term of operator and scan-rescan

Downloaded from http://ahajournals.org by on September 24, 2019

Table 1. Study Participant Characteristics

Patient Characteristics	Myocardial Infarction	Left Ventricular Hypertrophy	Cardiomyopathy	Other Pathology	Healthy Volunteers	*P* Value
No.	32	17	17	14	30	
Male	26 (81%)	14 (82%)	9 (53%)	5 (36%)	22 (73%)	-
BSA, m^2	1.94±0.4	1.99±0.4	2.00±0.4	1.88±0.5	1.82±0.5	-
Age, y	60±11	60±12	49±13	50±15	31±9	<0.0001
LVMi, g per m^2	77±22	92±19	93±33	61±16	61±10	<0.0001
EDVi, mL per m^2	75±22	73±15	122±45	76±23	88±13	<0.0001
ESVi, mL per m^2	34±17	24±7	72±51	28±13	33±9	<0.0001
SVi, mL per m^2	41±9	49±12	49±12	48±12	55±7	<0.0001
EF, %	56±8	68±8	45±17	64±7	63±5	<0.0001

Patients with left ventricular hypertrophy had diagnoses of hypertrophic cardiomyopathy (n=11), aortic stenosis (n=3), hypertensive heart disease (n=2), Anderson-Fabry disease (n=1). Patients with other cardiomyopathies had diagnoses of dilated cardiomyopathy (n=13), arrhythmogenic right ventricular cardiomyopathy (n=1), and left ventricular noncompaction (n=3). Other pathologies include chronic renal failure (n=5) and patients under cardio-oncology follow-up (n=9). Groups were compared using one-way ANOVA. BSA indicates body surface area; EDVi, (indexed) end diastolic volume; EF, ejection fraction; ESVi, (indexed) end systolic volume; LVMi, (indexed) left ventricular mass; and SVi, (indexed) stroke volume.

category. To assess the effect of LV impairment, a fixed effect of Scan 1 LVEF was included. Random effects included study subject in all models and operator when assessing training and effect of semi-automated technique.

Sample size required to detect a clinical change was calculated from the standardized difference (*d*) in each LV metric with a power of 90% and α of 0.05, where *d* is the desired clinical change divided by the SD of scan-rescan differences. The SE of measurement was calculated as the square root of the mean squared error obtained from one-way ANOVA. The minimal detectable change between 2 scans considered to be different was calculated as 2×SE of measurement. All tests were 2-tailed, and $P<0.05$ was considered statistically significant.

RESULTS

Scan-Rescan Cohort

The final data set of 110 scan-rescans represented patients with myocardial infarction (n=32), left ventricular hypertrophy (n=17, including hypertrophic cardiomyopathy), cardiomyopathy (n=17, dilated, arrhythmogenic right ventricular and left ventricular noncompaction cardiomyopathies), other pathology (n=14, cardio-oncology follow-up and chronic kidney disease), and healthy volunteers (n=30). All LV metrics differed significantly between diagnostic sub-groups, Table 1. One hundred six rescans (96%) were performed within one week (82% on the same day); 4 scans in healthy volunteers were performed between 1 week and 3 months.

Impact of Initial Training on Trainee-Expert Agreement

Training and implementation of a standard operating procedure significantly improved agreement between trainees and an expert for all LV metrics, Table II in the Data Supplement.

Impact of Techniques on Human Accuracy (Bias) and Precision

The thresholding and semi-automated techniques did not show a difference in accuracy with reference to manual contouring. As expected, the other 2 techniques showed over/under-estimation in EDV, ESV, and LVM; but not LVEF or SV: the click-draw technique measured the LV as larger and showed a trend to lower LVM (EDV 6.0±2.9 mL, $P=0.0449$; ESV 5.0±2.0 mL, $P=0.0133$; and LVM −6.7±3.6 g, $P=0.0671$); the level-set technique measured the LV as smaller and LVM as higher (EDV −8.3±2.9 mL, $P=0.0056$; ESV −5.8±2.0 mL, $P=0.0042$); LVM 15±4.3 g, $P<0.0004$).

For precision, however, there was no significant difference between techniques for either observer, Table III in the Data Supplement. This included both techniques that included papillary muscles in the blood-pool and those that included them within the myocardial mass, Figure 2. Given the similar precision between techniques, subsequent analysis of the complete data set by the expert and trainee used the thresholding technique, because it showed fewer large mistakes requiring manual correction.

Expert, Trainee and Automated Accuracy (Bias), Precision and Speed

There was good agreement between expert and trainee (intraclass correlation coefficients, 0.92–0.98) and automated analysis (intraclass correlation coefficients, 0.90–0.98) for all LV metrics, Figures I and II in the Data Supplement. Compared with expert analysis, trainee analysis measured the LV as slightly smaller, and LVEF and LVM as slightly higher. Automated analysis, conversely, measured the LV as slightly larger, and LVEF and LVM as slightly lower, Table IV in the Data Supplement.

Downloaded from http://ahajournals.org by on September 24, 2019

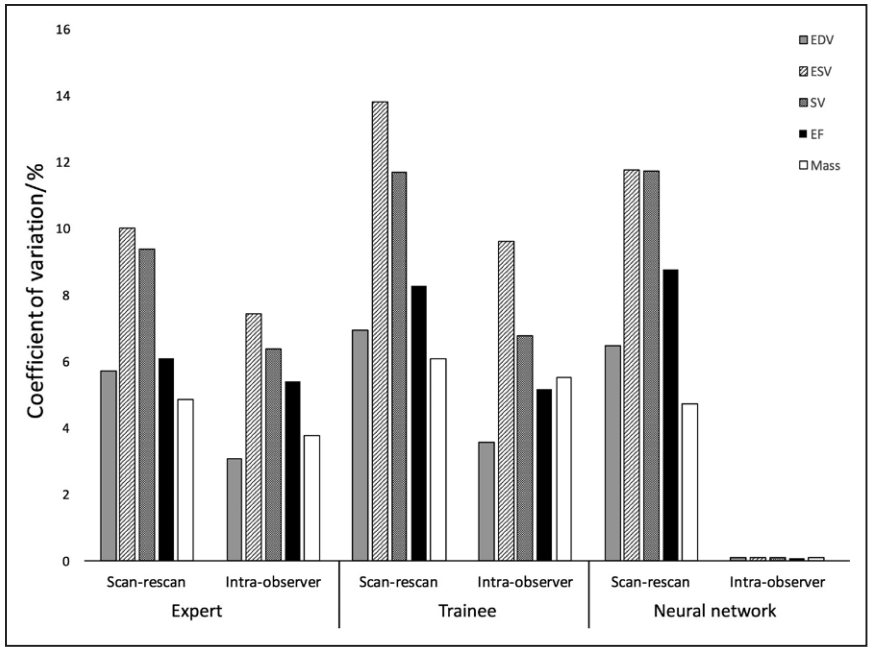

Figure 3. Scan-rescan coefficient of variation for expert, trainee, and automated neural network analysis.
All comparisons are not significant (Table V in the Data Supplement). EDV indicates left ventricular end diastolic volume; EF, left ventricular ejection fraction ESV, left ventricular end systolic volume; and SV, left ventricular stroke volume.

For precision, however, there was no significant difference between expert, trainee and automated analysis for all LV metrics, Table V in the Data Supplement. Table 2 and Figure 3 detail scan-rescan and intraobserver differences, Bland-Altman limits of agreement, and CV.

Human analysis time was 13 (interquartile range, 9–19) minutes per scan. Automated analysis time of a 25-phase short-axis stack was ≈0.07 minutes.

Sources of Error for Human and Automated Techniques

Average human CV for intraobserver, inter-observer and scan-rescan reproducibility were, for LVEF: 5.3%, 6.3%, and 7.2%, and for LVM: 4.6%, 7.6%, and 5.5% respectively, Figure 4 and Figure 5. For all LV metrics, human intraobserver CV was greater than half of the scan-rescan CV. For humans, scan-rescan CV was the greatest source of error for EDV, ESV, SV, and LVEF, while inter-observer error was the greatest source of error for LVM, Table VI in the Data Supplement. For automated analysis, there was zero intraobserver error as this technique was nonstochastic, that is, with the same image, the network would always generate an identical result. For all observers, precision was not influenced by the degree of impairment in LVEF.

Sample Size Estimates and Minimal Detectable Change

Calculation of sample size from these data shows that CMR requires 28 patients to detect a 3% change in LVEF; 12 patients to detect a 10 g change in LVM; and 17, 10, and 16 patients to detect a 10 mL change in EDV, ESV, and SV, respectively, Table 3. The percentage change in sample size of an expert for all LV metrics was similar to a trainee (×1.2–1.5) and automated analysis (×0.8–1.5). Sample size requirements were largest for patients with left ventricular hypertrophy.

For an individual patient, the minimal detectable change was 8.7% in LVEF or 20 g in LVM, based on expert analysis (with no difference when compared with automated analysis).

DISCUSSION

Despite reliance on measurements of LVEF and LVM for clinical decision-making and as end points in research studies, analysis is often not standardized and the relative contributions of error sources are imperfectly known. These data show that using current standardized image acquisition and multicenter, multi-vendor, multi-field strength, multi-disease, scan-rescan data at scale, measurement error was largely due to inconsistency in the human observer rather than variation in modifiable factors- clinician experience, scan acquisition, or human contour strategy (here performed using 5 techniques). This study also demonstrated for the first time that an automated analysis technique using deep learning has equivalent precision (scan-rescan reproducibility) to an expert, and yielded ≈13 minutes time-saving per scan, tested head-to-head on variable pathologies from multiple institutions. Clinicians can be confident in detecting a 9% change in LVEF or a 20 g change in LVM, this was similar if using an automated ML technique. Because the resource has the potential to test superiority of automated over human analysis,

Downloaded from http://ahajournals.org by on September 24, 2019

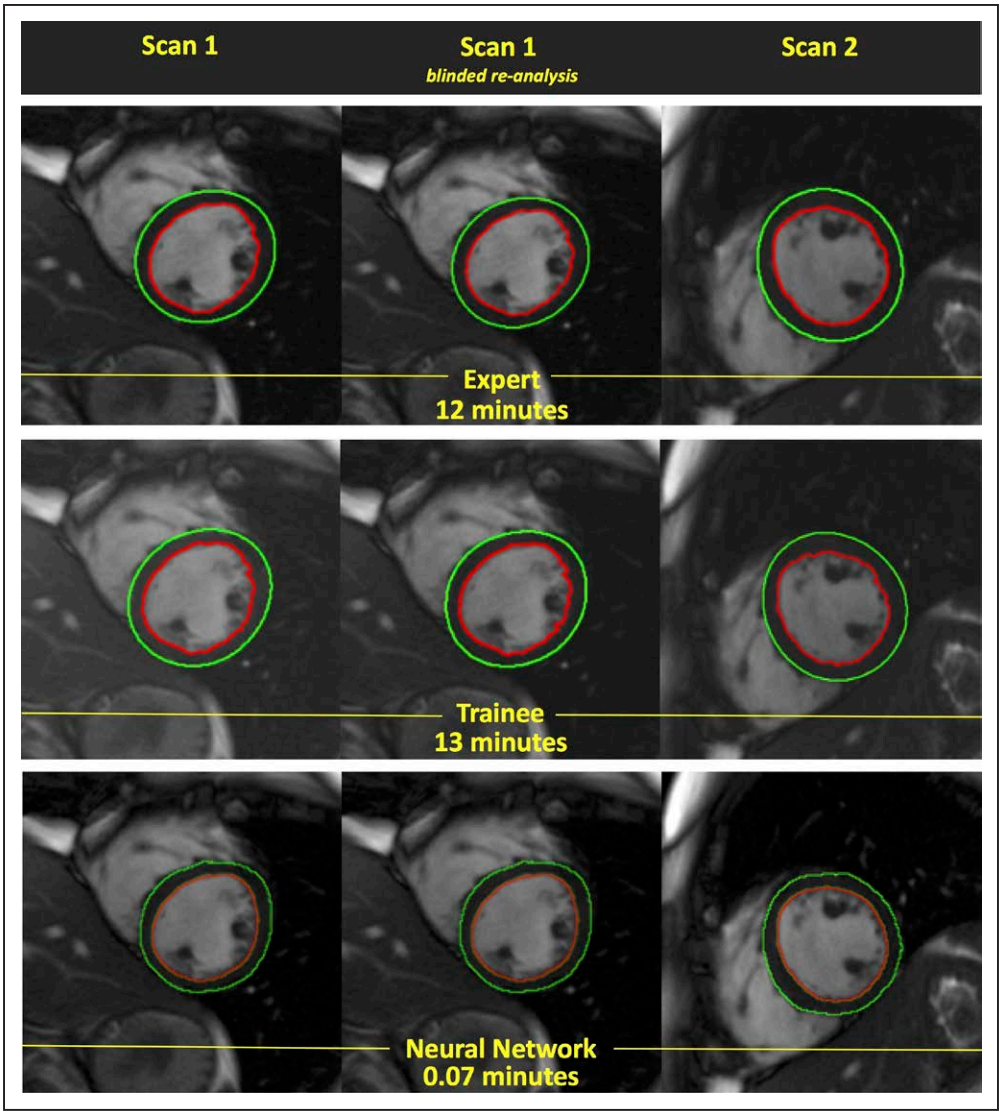

Figure 4. Examples of different contours for one cardio-oncology patient.
Analysis is by 3 observers (rows) over 3 data sets (columns), with average analysis timing per scan per observer reported. Left column: original scan 1, middle column: blinded scan 1 re-analysis; right column: repeat scan 2. Note this is one phase of one slice of ≈10 in each short-axis stack; that all human contours differ whereas the automated neural network scan 1 blinded reanalysis contours are identical. Note also the similar position but different piloting and orientation of the repeat scan 2.

scan-rescan data and training videos/standard operating procedures have been made available.

Previous studies investigating LV analysis by CMR have largely focused on inter-observer and intraobserver differences, usually using data sets from healthy individuals in single centers.[5,6] Clinical practice relies on scan-rescan precision, and includes sources of variation not previously captured in most data sets such as the scan acquisition, differences between institution, and disease states. This study looked at the efficacy of a number of potential strategies (decided a priori) that might reduce or maintain variability with time-saving through automation: clinician experience, training, human contouring methods, and deployment of an automated neural network segmentation approach.

Benefits of Automated Analysis

The adoption of ML can offer comparable precision with clinicians, with the time saving and global standardization that would ensue. Training of junior clinicians required a month-long program, compared with ≈9 hours for a neural network. Once trained, clinicians required an average of 13 minutes for analysis per scan, compared with ≈4 seconds for a neural network. In the UK, an estimated 2275 scans per million adults are needed annually, performed in 61 centers.[20] Automating this one aspect of CMR analysis alone would, therefore, potentially translate into a saving of 54 clinician-days per center. Accurate automated segmentation is a bridge to reliable extraction of more information from the same imaging beyond

Downloaded from http://ahajournals.org by on September 24, 2019

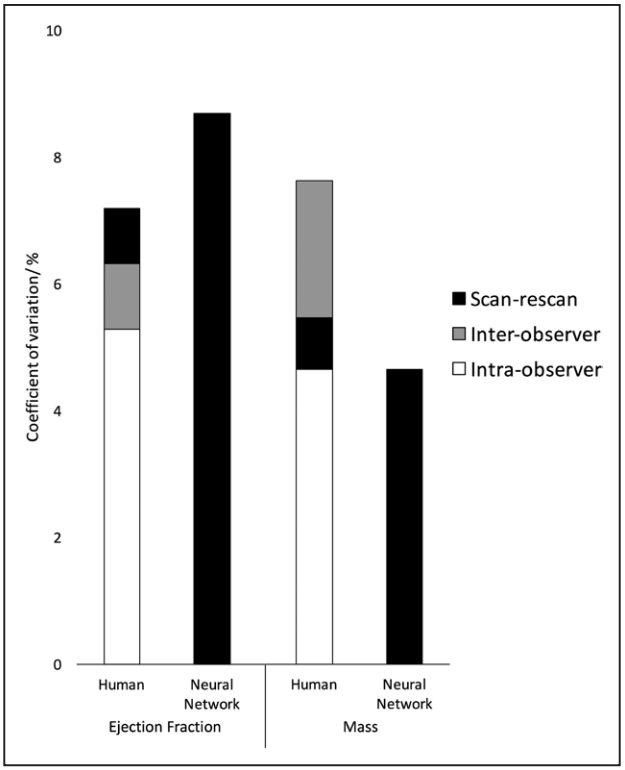

Figure 5. Contributions of intra-, inter-observer, and scan-rescan error to human measurement variability.
For ejection fraction, human error (coefficient of variation) is incremental from intraobserver, inter-observer (on the same scan), to inter-scan. For mass, human inter-observer error is greatest. Scan-rescan error is similar for human and automated neural network analysis, however the majority of error for human analysis is related to the observer suggesting that automated techniques have other sources that can be addressed to surpass human performance. Human error is the average of both human observers.

established imaging biomarkers. In combination with time saving, this maximizes use of acquired data in a value-based manner.

ML Techniques can Surpass Human Precision

Given that the greatest sources of measurement error were human factors (ie, nonmodifiable intra- and inter-observer variability), we believe that, with improvement, it is only a matter of time before automated approaches are super-human, with cascading consequences in clinical (confidence, smallest detectable difference) and research (trial size) domains of increased precision. Improvement could be related to either training data or the network itself, and comparison of scan-rescan precision against that of an expert offers the ability to show superior performance, and provides an important validation step towards real-world scalability.

Similar performance of automated techniques to humans has previously been shown by comparing the degree of inter-observer agreement between 2 clinicians and clinician-network agreement in typically healthy subjects.[10,11] Evaluation of new automated segmentation techniques is typically performed against expert ground truth in medical imaging grand challenge data sets, and presented as contour comparisons (such as mean error or dice index).[21,22] Expert segmentation for the ground truth comparisons is time consuming meaning cases numbers are necessarily small, limited to a few centers and pathologies, and such metrics are not intuitive for clinicians.[6] UK Biobank data and the MICCAI ACDC 2017 challenge have addressed issues of cohort size and differences between scanners, but using this approach it is not possible to identify a technique that is superior to human analysis, despite growing appreciation of the high intra- and inter-observer variability in human measurements.[6,10,11] Comparison of measurement precision between techniques using test:retest methodology avoids this limitation, and this cohort enables identification of methods that are both generalizable and superior to humans.

Superior performance will require potentially larger and more variable pathology datasets facilitated by adversarial training, or transfer learning.[23,24] However, if training is performed by one expert annotating each dataset, neural networks will be trained to minimize between subject differences, but not differences between or within observers. Training on repeated measures may minimize these errors. This could also be surmounted by the use of stronger priors,[25] 3D neural networks or even limiting the reliance on annotation through deeper, more intelligent pixel classification.

Neural network approaches, however, do show limitations. Data must be standardized before analysis, and neural networks are computationally expensive to train and require clinician-facing interfaces before widespread implementation. Biologically implausible segmentations are also possible and therefore results require human review.[11]

Sample Size Estimates

This study provides benchmark precision metrics that reflect a range of pathologies and institutions. Required sample sizes to detect a standardized difference was greatest for patients with left ventricular hypertrophy who had increased LVM and small systolic cavity volumes. The sample size required to detect a clinically important change was between 10 and 28 patients for different LV metrics, which is higher than previous estimates.[5,12,13] This data set however is different due to its high variability, and should these results be considered for future study design, there may be reasons that these results either over- or under-represent anticipated performance for a specific real-world task. Factors that make precision higher here include: excellent training/

Downloaded from http://ahajournals.org by on September 24, 2019

Table 2. Scan-Rescan and Intraobserver Reproducibility Stratified by Observer

	Scan-Rescan Reproducibility					Intraobserver Reproducibility				
	EDV, mL	ESV, mL	SV, mL	EF (%)	Mass, g	EDV, mL	ESV, mL	SV, mL	EF (%)	Mass, g
Expert										
Mean	159±51	69±46	90±23	59±12	142±44	160±51	69±45	91±24	59±12	143±44
Difference	9±8	6±6	9±8	4±3	7±7	5±4	5±5	6±5	3±3	4±3
CV/%	5.7 (4.7–6.8)	10.0 (8.1–11.8)	9.4 (7.8–11.0)	6.1 (5.2–7.1)	4.8 (4.1–5.6)	3.1 (2.5–3.7)	7.4 (6.1–8.9)	6.4 (5.1–7.8)	5.4 (3.9–6.9)	3.8 (3.25–4.29)
BA limits	−23 to 26	−18 to 17	−20 to 25	−9 to 10	−19 to 19	−12.5 to 12.2	−13 to 13	−16 to 15	−8 to 8	−16 to 15
Trainee										
Mean	156±51	64±47	92±26	61±14	146±45	157±51	64±45	92±26	61±13	144±43
Difference	10±9	7±7	1 0±9	5±4	9±8	5±6	5±5	6±6	3±3	8±7
CV/%	7.0 (5.8–8.2)	13.8 (11.4–16.2)	11.7 (9.4–13.9)	8.3 (5.6–10.3)	6.1 (5.2–6.9)	3.6 (2.9–4.3)	9.6 (7.4–11.8)	6.8 (5.5–8.2)	5.2 (4.2–6.2)	5.5 (4.2–6.7)
BA limits	−25 to 29	−19 to 20	−25 to 28	−12 to 12	−25 to 23	−15 to 16	−14 to 14	−17 to 18	−8 to 9	−22 to 20
Automated										
Mean	166±53	77±47	89±26	56±12	135±40	–	–	–	–	–
Difference	10±9	8±8	10±9	4±4	6±6	–	–	–	–	–
CV/ %	6.5 (5.2–7.8)	11.8 (8.5–14.6)	11.7 (9.1–14.1)	8.8 (6.1–11.1)	4.7 (4.0–5.6)	–	–	–	–	–
BA limits	−25 to 28	−22 to 20	−24 to 29	−11 to 12	−17 to 16	–	–	–	–	–

Data are presented as cohort mean±SD; absolute difference between scans±SD; CV and 95% CI; and BA limits. BA limits indicates Bland-Altman limits of agreement; CV, within-subject coefficient of variation; EDV, end diastolic volume; EF, ejection fraction; ESV, end systolic volume; and SV, stroke volume.

standardization and operator selection; bias from using only the best 110 studies from 118 with (by definition) no further dropout; scan-rescan at short time intervals; a few expert centers only. Factors that make precision lower here include: inclusion of multiple diseases; multiple centers; multiple scanners and blinded analysis. While follow-up studies may be analyzed consecutively in clinical practice, we ensured that the 2 scans acquired for each subject were assigned different randomized study numbers to minimize bias in this study. There was no review of study data at study completion, with all analyzed data sets included in the results.

Sources of Human Measurement Error

It is generally accepted that there is incremental variation from intraobserver, inter-observer and scan-rescan (physiological and technical) differences.[26] These data

Table 3. Sample Size Estimates Stratified by Observer and Pathology

Sample Size Estimates (*n*), α=0.05, 90% Power										
	EDV		ESV		SV		EF		Mass	
	10 mL change		10 mL change		10 mL change		3% change		10g change	
	SD	n	SD	n	SD	n	SD	n	SD	n
Whole cohort										
Expert	12	17	9	10	11	16	5	28	10	12
Trainee	14	×1.3	10	×1.2	14	×1.3	6	×1.5	12	×1.5
Automated	13	×1.2	11	×1.4	13	×1.3	6	×1.5	8	×0.8
Sub-groups (expert only)										
MI	12	18	7	7	11	15	5	31	11	14
LVH	13	20	9	11	13	20	5	36	8	9
CM	12	17	10	12	9	10	4	19	13	20
Other pathology	14	23	6	6	12	17	3	15	6	7
HV	9	11	10	13	11	16	5	26	9	11

For Trainee and Automated neural network analysis, sample size is represented as a proportional change to Expert analysis. Sample size estimates stratified by pathology are presented for expert analysis only.

CM indicates cardiomyopathy; EDV, end diastolic volume; EF, ejection fraction; ESV, end systolic volume; HV, healthy volunteers; LVH, left ventricular hypertrophy; MI, myocardial infarction; SD, standard deviation; and SV, stroke volume.

Downloaded from http://ahajournals.org by on September 24, 2019

demonstrated that human (intraobserver) error (CV) was greater than half of scan-rescan error, an effect that was not minimized by an expert when compared with junior clinicians after appropriate training, despite 15 years' additional experience. A training program combined with standard operating procedures was an effective approach to improve inter-observer agreement, by standardizing basal blood volume and papillary muscles, as previously reported.[7,8] Semi-automated techniques, including a novel level-set approach with minimal user interaction also did not improve precision over manual techniques, a finding replicated in 2 observers. However, human contour strategies resulted in potentially clinically relevant differences for an individual patient, emphasizing the need to interpret reference ranges in the context of the technique from which they were derived.

To improve human measurement precision, an improved focus on unifying a systematic approach to analysis and greater acquisition standardization appears important. This would require investigating piloting of the short-axis cine stack, loading conditions, or sequence improvements that improve myocardial contrast with epicardial fat.[27]

Study Limitations

We measured variability using a relatively small number of observers, however by including both trainees and an international expert as the gold standard observer, the errors measured are likely to be representative. Both scans for each patient were acquired in the same institution using the same protocol, and therefore we have not assessed scan-rescan precision between institutions. The scan-rescan interval was short (with 82% of studies acquired on the same day), and we are, therefore, unable to assess the contribution of physiological variability across months or years. The precision of right ventricular assessment was not within the scope of this study. We analyzed the performance of a ML approach but not its prospective clinical application.

CONCLUSIONS

Automated ML techniques for LV analysis match human precision and perform substantially faster. Based on multicenter, multi-vendor, multi-field strength, multi-disease data, a 9% change in ejection fraction can be detected confidently by expert clinicians, and this is similar using automated analysis. Given that a major source of measurement variability is attributable to the observer, automated approaches offer the future potential to surpass human experts, demonstrable using this scan-rescan resource.

Downloaded from http://ahajournals.org by on September 24, 2019

ARTICLE INFORMATION

Received April 3, 2019; accepted July 25, 2019.

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCIMAGING.119.009214.

Authors

Anish N. Bhuva, MBBS; Wenjia Bai, PhD; Clement Lau, MBChB; Rhodri H. Davies, PhD; Yang Ye, PhD; Heeraj Bulluck, PhD; Elisa McAlindon, PhD; Veronica Culotta, MBBS; Peter P. Swoboda, PhD; Gabriella Captur, PhD; Thomas A. Treibel, PhD; Joao B. Augusto, MD; Kristopher D. Knott, MBBS; Andreas Seraphim, MBBS; Graham D. Cole, PhD; Steffen E. Petersen, PhD; Nicola C. Edwards, PhD; John P. Greenwood, PhD; Chiara Bucciarelli-Ducci, PhD; Alun D. Hughes, PhD; Daniel Rueckert, PhD; James C. Moon, MD; Charlotte H Manisty, PhD

Correspondence

Charlotte H. Manisty, PhD, Department of Cardiac Imaging, Barts Heart Centre, W Smithfield, London EC1A 7BE. Email c.manisty@ucl.ac.uk

Affiliations

Institute for Cardiovascular Science, University College London, United Kingdom (A.N.B., H.B., G.C., T.A.T., J.B.A., K.D.K., A.S., A.D.H., J.C.M., C.H.M.). Department of Cardiovascular Imaging, Barts Heart Centre, Barts Health NHS Trust, London, United Kingdom (A.N.B., C.L., R.H.D., Y.Y., V.C., G.C., T.A.T., J.B.A., K.D.K., A.S., S.E.P., J.C.M., C.H.M.). Data Science Institute and Department of Medicine (W.B.), Department of Computing (D.R.), Imperial College London, South Kensington Campus, United Kingdom. William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, United Kingdom (C.L., S.E.P.). Department of Cardiology, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, People's Republic of China (Y.Y.). Bristol Heart Institute, Bristol NIHR Biomedical Research Centre, University Hospitals Bristol NHS Trust and University of Bristol, United Kingdom (E.M., C.B.-D.). Heart and Lung Centre, New Cross Hospital, Wolverhampton, United Kingdom (E.M.). Multidisciplinary Cardiovascular Research Centre and Division of Biomedical Imaging, Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, United Kingdom (P.P.S., J.P.G.). Imperial College London, National Heart and Lung Institute, Hammersmith Hospital, United Kingdom (G.D.C.). Auckland City Hospital, New Zealand and Institute of Cardiovascular Science, University of Birmingham (N.C.E.).

Acknowledgments

The authors acknowledge all research staff at each site for their contribution to recruitment. They also acknowledge the contribution of the British Society for Cardiovascular Magnetic Resonance Valve Consortium for developing, acquiring and making available the training data set segmentations.

Sources of Funding

This study was supported by Barts Charity (MGU0302). A.N. Bhuva is supported by a doctoral research fellowship from the British Heart Foundation (FS/16/46/32187). Drs Moon and Manisty are directly and indirectly supported by the University College London Hospitals and Drs Moon, Manisty, and Petersen by the Barts Biomedical Research Centre. A.N. Bhuva received support from the British Heart Foundation (PG/13/6/29934), the National Institute for Health Research University College London Hospitals Biomedical Research Centre, and works in a unit that receives support from the UK Medical Research Council (MC_UU_12019/1). CBD is in part supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. Drs Bai and Petersen acknowledge support from the SmartHeart EPSRC program grant (EP/P001009/1).

Disclosures

Dr Bucciarelli-Ducci is a consultant for Circle Cardiovascular Imaging (Calgary, Canada). The other authors report no conflicts.

REFERENCES

1. Multicenter Postinfarction Research Group. Risk stratification and survival after myocardial infarction. *N Engl J Med*. 1983;309:331–336.

2. Bluemke DA, Kronmal RA, Lima JA, Liu K, Olson J, Burke GL, Folsom AR. The relationship of left ventricular mass and geometry to incident cardiovascular events: the MESA (Multi-Ethnic Study of Atherosclerosis) study. *J Am Coll Cardiol*. 2008;52:2148–2155. doi: 10.1016/j.jacc.2008.09.014
3. Marwick TH. Ejection fraction pros and cons: JACC state-of-the-art review. *J Am Coll Cardiol*. 2018;72:2360–2379. doi: 10.1016/j.jacc.2018.08.2162
4. Kramer CM, Barkhausen J, Flamm SD, Kim RJ, Nagel E; Society for Cardiovascular Magnetic Resonance Board of Trustees Task Force on Standardized Protocols. Standardized cardiovascular magnetic resonance (CMR) protocols 2013 update. *J Cardiovasc Magn Reson*. 2013;15:91. doi: 10.1186/1532-429X-15-91
5. Bellenger NG, Davies LC, Francis JM, Coats AJ, Pennell DJ. Reduction in sample size for studies of remodeling in heart failure by the use of cardiovascular magnetic resonance. *J Cardiovasc Magn Reson*. 2000;2:271–278. doi: 10.3109/10976640009148691
6. Suinesiaputra A, Bluemke DA, Cowan BR, Friedrich MG, Kramer CM, Kwong R, Plein S, Schulz-Menger J, Westenberg JJ, Young AA, Nagel E. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J Cardiovasc Magn Reson*. 2015;17:63. doi: 10.1186/s12968-015-0170-9
7. Miller CA, Jordan P, Borg A, Argyle R, Clark D, Pearce K, Schmitt M. Quantification of left ventricular indices from SSFP cine imaging: impact of real-world variability in analysis methodology and utility of geometric modeling. *J Magn Reson Imaging*. 2013;37:1213–1222. doi: 10.1002/jmri.23892
8. 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;9:777–783. doi: 10.1080/10976640701545073
9. Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. *Med Image Anal*. 2011;15:169–184. doi: 10.1016/j.media.2010.12.004
10. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi M, Zemrak F, Fung K, Paiva JM, Carapella V, Kim YJ, Suzuki H, Kainz B, Matthews PM, Petersen SE, Piechnik SK, Neubauer S, Glocker B, Rueckert D. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*. 2017;20:1–12. doi: 10.1186/s12968-018-0471-x
11. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, Cetin I, Lekadir K, Camara O, Balleser M. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging*. 2018;62:1–12. doi: 10.1109/TMI.2018.2837502
12. Grothues F, Smith GC, Moon JC, Bellenger NG, Collins P, Klein HU, Pennell DJ. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *Am J Cardiol*. 2002;90:29–34. doi: 10.1016/s0002-9149(02)02381-0
13. Moody WE, Edwards NC, Chue CD, Taylor RJ, Ferro CJ, Townend JN, Steeds RP. Variability in cardiac MR measurement of left ventricular ejection fraction, volumes and mass in healthy adults: defining a significant change at 1 year. *Br J Radiol*. 2015;88:20140831. doi: 10.1259/bjr.20140831
14. Schulz-Menger J, Bluemke DA, Bremerich J, Flamm SD, Fogel MA, Friedrich MG, Kim RJ, von Knobelsdorff-Brenkenhoff F, Kramer CM, Pennell DJ, Plein S, Nagel E. 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;15:35. doi: 10.1186/1532-429X-15-35
15. Captur G, Radenkovic D, Li C, Liu Y, Aung N, Zemrak F, Tobon-Gomez C, Gao X, Elliott PM, Petersen SE, Bluemke DA, Friedrich MG, Moon JC. Community delivery of semiautomated fractal analysis tool in cardiac mr for trabecular phenotyping. *J Magn Reson Imaging*. 2017;46:1082–1088. doi: 10.1002/jmri.25644
16. Musa TA, Treibel TA, Vassiliou VS, Captur G, Singh A, Chin C, Dobson LE, Pica S, Loudon M, Malley T, Rigolli M, Foley JRJ, Bijsterveld P, Law GR, Dweck MR, Myerson SG, McCann GP, Prasad SK, Moon JC, Greenwood JP. Myocardial scar and mortality in severe aortic stenosis. *Circulation*. 2018;138:1935–1947. doi: 10.1161/CIRCULATIONAHA.117.032839
17. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255–268.
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
19. Najjar SS, Scuteri A, Shetty V, Wright JG, Muller DC, Fleg JL, Spurgeon HP, Ferrucci L, Lakatta EG. Pulse wave velocity is an independent predictor of the longitudinal increase in systolic blood pressure and of incident hypertension in the Baltimore Longitudinal Study of Aging. *J Am Coll Cardiol*. 2008;51:1377–1383. doi: 10.1016/j.jacc.2007.10.065
20. NHS England. Standard Contract for CMR. 2013. https://www.england.nhs.uk. Accessed January 18, 2019.
21. Radau P, Lu Y, Connelly K, Paul G, Dick AJ; Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI. The MIDAS Journal- Cardiac MR Left Ventricle Segmentation Challenge 2009, http://hdl.handle.net/10380/3070. Accessed September 15, 2016.
22. Suinesiaputra A, Cowan BR, Al-Agamy AO, Elattar MA, Ayache N, Fahmy AS, Khalifa AM, Medrano-Gracia P, Jolly MP, Kadish AH, Lee DC, Margeta J, Warfield SK, Young AA. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. *Med Image Anal*. 2014;18:50–62. doi: 10.1016/j.media.2013.09.001
23. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. *Circ Cardiovasc Imaging*. 2017;10:1–10. doi: 10.1161/CIRCIMAGING.117.005614
24. Tao Q, Yan W, Wang Y, Paiman E, Shamonin D, Garg P, Plein S, Huang L, Xia L, Sramko M, Tintera J, de Roos A, Lamb HJ, van der Geest RJ. Deep Learning – based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, Multicenter Study. *Radiology*. 2018;290:81–88. doi: 10.1148/radiol.2018180513
25. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal*. 2016;30:108–119. doi: 10.1016/j.media.2016.01.005
26. Bellenger NG, Marcus NJ, Rajappan K, Yacoub M, Banner NR, Pennell DJ. Comparison of techniques for the measurement of left ventricular function following cardiac transplantation. *J Cardiovasc Magn Reson*. 2002;4:255–263. doi: 10.1081/jcmr-120003951
27. Marchesseau S, Ho JX, Totman JJ. Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: a reproducibility study. *Eur J Radiol Open*. 2016;3:60–66. doi: 10.1016/j.ejro.2016.03.003

Downloaded from http://ahajournals.org by on September 24, 2019