Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study

Jasper Tromp*, Paul J Seekings*, Chung-Lieh Hung, Mathias Bøtcher Iversen, Matthew James Frost, Wouter Ouwerkerk, Zhubo Jiang, Frank Eisenhaber, Rick S M Goh, Heng Zhao, Weimin Huang, Lieng-Hsi Ling, David Sim, Patrick Cozzone, A Mark Richards, Hwee Kuan Lee, Scott D Solomon, Carolyn S P Lam, Justin A Ezekowitz

Lancet Digit Health 2022; 4: e46–54

Published Online December 1, 2021 https://doi.org/10.1016/S2589-7500(21)00235-1

*Contributed equally

National Heart Centre Singapore, Singapore (J Tromp MD, W Ouwerkerk PhD, D Sim MBBS, Prof C S P Lam MBBS); **Duke-NUS Medical School, Singapore** (J Tromp, Prof C S P Lam); **Saw Swee Hock School of Public Health, National University of Singapore & National University Health System, Singapore** (J Tromp); **Bioinformatics Institute** (P J Seekings PhD, F Eisenhaber PhD, H K Lee PhD), **Genome Institute of Singapore** (F Eisenhaber), **Institute of High Performance Computing** (R S M Goh PhD, H Zhao PhD), **Institute for Infocomm Research** (W Huang PhD), **and Singapore Bioimaging Consortium, Biomedical Sciences Institutes** (Prof P Cozzone PhD), **Agency for Science, Technology and Research (A*STAR), Singapore; Us2.ai, Singapore** (P J Seekings, M B Iversen BSc, M J Frost, Z Jiang MSc); **Department of Medicine and Institute of Biomedical Sciences, Mackay Medical College, Taipei, Taiwan** (C-L Hung MD); **Cardiovascular Division, Department of Internal Medicine, Mackay Memorial Hospital, Taipei, Taiwan** (C-L Hung); **Department of Dermatology, Amsterdam UMC, University of Amsterdam, Amsterdam Infection and Immunity Institute, Amsterdam, Netherlands** (W Ouwerkerk); **School of Biological Science, Nanyang Technological University, Singapore** (F Eisenhaber); **National University Heart Centre, Singapore** (L-H Ling MBBS, Prof A M Richards MBChB); **Yong**

Abstract

Background Echocardiography is the diagnostic modality for assessing cardiac systolic and diastolic function to diagnose and manage heart failure. However, manual interpretation of echocardiograms can be time consuming and subject to human error. Therefore, we developed a fully automated deep learning workflow to classify, segment, and annotate two-dimensional (2D) videos and Doppler modalities in echocardiograms.

Methods We developed the workflow using a training dataset of 1145 echocardiograms and an internal test set of 406 echocardiograms from the prospective heart failure research platform (Asian Network for Translational Research and Cardiovascular Trials; ATTRaCT) in Asia, with previous manual tracings by expert sonographers. We validated the workflow against manual measurements in a curated dataset from Canada (Alberta Heart Failure Etiology and Analysis Research Team; HEART; n=1029 echocardiograms), a real-world dataset from Taiwan (n=31 241), the US-based EchoNet-Dynamic dataset (n=10 030), and in an independent prospective assessment of the Asian (ATTRaCT) and Canadian (Alberta HEART) datasets (n=142) with repeated independent measurements by two expert sonographers.

Findings In the ATTRaCT test set, the automated workflow classified 2D videos and Doppler modalities with accuracies (number of correct predictions divided by the total number of predictions) ranging from 0·91 to 0·99. Segmentations of the left ventricle and left atrium were accurate, with a mean Dice similarity coefficient greater than 93% for all. In the external datasets (n=1029 to 10 030 echocardiograms used as input), automated measurements showed good agreement with locally measured values, with a mean absolute error range of 9–25 mL for left ventricular volumes, 6–10% for left ventricular ejection fraction (LVEF), and 1·8–2·2 for the ratio of the mitral inflow E wave to the tissue Doppler e' wave (E/e' ratio); and reliably classified systolic dysfunction (LVEF <40%, area under the receiver operating characteristic curve [AUC] range 0·90–0·92) and diastolic dysfunction (E/e' ratio ≥13, AUC range 0·91–0·91), with narrow 95% CIs for AUC values. Independent prospective evaluation confirmed less variance of automated compared with human expert measurements, with all individual equivalence coefficients being less than 0 for all measurements.

Interpretation Deep learning algorithms can automatically annotate 2D videos and Doppler modalities with similar accuracy to manual measurements by expert sonographers. Use of an automated workflow might accelerate access, improve quality, and reduce costs in diagnosing and managing heart failure globally.

Funding A*STAR Biomedical Research Council and A*STAR Exploit Technologies.

Introduction

Heart failure is a significant public health problem worldwide.[1] Early diagnosis and treatment can prevent disease progression and reduce the burden on health-care systems. Echocardiography is the most commonly used cardiac imaging modality and is generally considered the primary method for assessing cardiac structure and function in the diagnosis of heart failure.[2–4] Although echocardiography is non-invasive, safe, and highly acceptable to patients, cases that require trained specialists to interpret echocardiograms can limit accessibility.[2,5]

Advances in deep learning have made automated analysis of medical images possible.[6,7] Previous attempts to automate echocardiogram interpretation have focused solely on view identification[8,9] or the quantification of systolic function,[10–13] or have not included external validation in patients with abnormal findings.[10,12,13] However, more than half of patients with heart failure have either a mid-range ejection fraction or preserved ejection fraction (HFpEF), and assessment of diastolic function is of crucial importance across cardiac disease states.[3] The diagnosis of HFpEF is challenging, and in parallel to clinical evaluation, relies on identifying structural and functional changes associated with increased left ventricular filling pressure.[3,14,15] Therefore, a fully automated workflow to assess systolic and diastolic function parameters is a crucial unmet need.

Loo Lin School of Medicine, National University of Singapore, Singapore (L-H Ling); Cardiovascular Research Institute, National University Health System, Singapore (Prof A M Richards); Christchurch Heart Institute, University of Otago, Christchurch, New Zealand (Prof A M Richards); Image and Pervasive Access Lab, CNRS UMI 2955, Singapore (H K Lee); Singapore Eye Research Institute, Singapore (H K Lee); Cardiovascular Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA (Prof S D Solomon MD); Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands (Prof C S P Lam); Canadian VIGOUR Centre, University of Alberta, Edmonton, AB, Canada (Prof J A Ezekowitz MBBCH)

Correspondence to:
Professor Justin A Ezekowitz,
Canadian VIGOUR Centre,
University of Alberta, Edmonton,
AB T6G 2E1, Canada
jae2@ualberta.ca

See Online for appendix

Research in context

Evidence before this study

We searched PubMed for papers published between Jan 1, 2000, and April 1, 2020, using the search terms "deep learning" OR "artificial intelligence" OR "machine learning" AND "echocardiography". We did not apply restrictions on study type or language. Previous attempts that used deep learning algorithms to classify and annotate echocardiograms focused either exclusively on classification without annotation or only assessed systolic function. Importantly, previous attempts commonly did not perform independent external validation.

Added value of this study

This study presents a fully automated deep learning-based workflow for automating the classification and annotation of echocardiographic videos. Measures of agreement between human measurements and automated measurements were high for systolic and diastolic dysfunction. Automated measurements were able to diagnose systolic and diastolic dysfunction with high ability (area under the receiver operating characteristic curve at or around 0·9). Importantly, results were validated in three independent cohorts from Canada, Taiwan, and the USA. A separate prospective assessment showed that the variability of automated measurements was smaller than the variability of manual measurements by expert sonographers.

Implications of all the available evidence

The present deep learning model annotated echocardiograms and showed the ability to detect systolic and diastolic dysfunction with similar accuracy and lower variability than human experts. These results highlight the potential of deep learning algorithms to help interpret echocardiograms. Ultimately, the use of automated workflows can democratise access and use of echocardiography in resource-limited settings.

To address this need, we developed a fully automated deep learning-based workflow to estimate parameters of cardiac systolic and diastolic function by echocardiography. First, we classified echocardiographic videos according to the most used views. Second, we used automatic annotators on the basis of convolutional neural networks (CNNs) to quantify cardiac chamber volumes, and left ventricular systolic function (ejection fraction) and diastolic function (ratio of the mitral inflow E wave to the tissue Doppler e' wave; E/e' ratio). Finally, we used the automated workflow to generate interpretable annotations and compared with human measurements in independent, sex-balanced, and ethnically diverse cohorts.

Methods

Study design and datasets

For more on ATTRACT see https://www.a-star.edu.sg/attract

We prototyped the automated deep learning-based workflow at key research institutes (Bioinformatics Institute, Institute of High Performance Computing, and Institute for Infocomm Research) of the Agency for Science, Technology and Research (A*STAR) of Singapore, using data from the Asian Network for Translational Research and Cardiovascular Trials (ATTRaCT) programme. The ATTRaCT platform contains data from the Asian Sudden Cardiac Death in Heart Failure registry and the ATTRaCT cohort, representing 11 countries (China, Hong Kong, India, Indonesia, Japan, Malaysia, Philippines, Singapore, South Korea, Taiwan, and Thailand). We used a total of 1145 individual echocardiograms from 1076 patients in the training set. The test set consisted of 406 separate echocardiograms from 406 patients. Sample sizes were based on the availability of data. We then validated the reliability of automated measurements in three external datasets without additional tuning. There was no overlap between patients in the test and validation sets. First, we validated the workflow using the Alberta Heart Failure Etiology and Analysis Research Team (HEART) study. The Alberta HEART study was a prospective observational study in Canada, with protocolised echocardiograms, read by two sonographers according to American Society of Echocardiography standards.[16] Baseline characteristics of the Canada cohort are shown in the appendix (p 1). In total, 621 participants had an echocardiogram performed at baseline and 408 participants had an echocardiogram performed at 1-year follow-up, giving a total of 1029 individual echocardiograms. Second, we validated the automated measurements in a large real-world dataset of 31 241 echocardiograms from 9289 individuals from Mackay Memorial Hospital, Taipei, Taiwan, which is one of the largest tertiary medical centres in Taiwan. The cohorts from Canada and Taiwan included both participants with heart failure and participants without heart failure. Baseline characteristics of the Taiwan cohort are shown in the appendix (p 2). Lastly, we validated the measurement of left ventricular ejection fraction (LVEF) in the US-based EchoNet-Dynamic dataset, which includes 10 030 echocardiograms with ground truth measurements of LVEF.[11] Manual measurements in the Taiwan and US cohorts were done by multiple expert sonographers. Institutional review boards of participating centres approved each cohort study. Patients in ATTRaCT and the Alberta HEART study signed informed consent allowing use of data in secondary studies. Data from Mackay Memorial Hospital was retrospectively identified. A waiver of consent was obtained from the Mackay Memorial Hospital institutional review board. The EchoNet-Dynamic data is a publicly available dataset of anonymised imaging studies from patients obtained during standard clinical care. A waiver of consent was obtained from the institutional review board of Stanford

University (Stanford, CA, USA), as detailed previously.[11] Complete information on inclusion and exclusion criteria of the different cohorts are provided in the appendix (pp 15–17). Individual echocardiograms did not always have all two-dimensional (2D) or Doppler modality views available or manual measurements available for comparison. Therefore, we report results for studies with available views and manually measured values.

Video-based deep learning models for view classification and annotation

The human workflow for performing echocardiographic measurements consist of: (1) identifying the correct 2D or Doppler modality view; (2) manually segmenting and drawing annotations on 2D videos to outline cardiac chambers, or annotating Doppler modality views for specific measurements; and (3) manually referring to standard reference guidelines to identify if any of the potentially dozens of measurements performed fall within the normal range of values for sex and age. The deep learning-based workflow automates the entire process by rapidly analysing the digital imaging and communications in medicine (DICOM) files of a patient's echocardiographic exam without the need for human intervention; from the classification of views to annotations, measurements, and generation of decision-support outcomes about cardiac structure and function (appendix p 10).

The first stage of the workflow categorises the imaging mode and view of each file into: 2D video, 2D video and colour Doppler, pulsed wave tissue Doppler imaging (PWTDI), M-mode, pulsed wave Doppler, and continuous wave Doppler. Videos were then parsed through two separate workflows for 2D videos and Doppler modalities (appendix p 10). In the second stage, 2D video and modalities were pre-processed and classified into views (appendix p 10). 2D videos were classified as apical 4 chamber (A4C) view, apical 2 chamber (A2C) view, parasternal long axis (PLAX) view, or 2D other views, and focused versions of the main views. Doppler modalities were classified as PWTDI (lateral, medial, and tricuspid regurgitation peak velocity [TrV]), M-mode (TrV and other views), pulsed wave (mitral valve and other views), continuous wave (TrV, aortic outflow velocity, and other views). DICOM images were pre-processed by cropping the echo sector to a tight square, or the velocity trace to a tight rectangle. The ECG was converted to a trace, to determine P, R, and T phases of the cardiac cycle if present.

2D videos were classified into views by one of two different classifiers. The first classifier was a supervised CNN, composed of four convolutional layers, a dense layer, and a softmax output layer. This model was trained with a categorical cross-entropy loss function. The second classifier was a modified version of an unsupervised deep clustering CNN[17] (appendix p 10), trained with mean squared error and Kullback–Leibler loss functions. The Doppler modalities view classifier consisted of an ensemble of CNN models trained with the echo or velocity trace images and the categorical ground truth labels, with a categorical cross entropy loss function and a softmax output layer (appendix p 10). We trained the models on 55 487 images from 1145 individual echocardiograms (appendix pp 2–3).

Classified views of 2D videos and Doppler modalities were annotated. Experienced sonographers annotated a total of 20 828 available images (appendix p 4) for training. A combination of CNNs were used for annotating 2D videos and doppler modalities. Segmentation models were based on a UNet style architecture with a sigmoid output layer and trained with the combined binary cross-entropy and Dice loss function. For A2C and A4C views the endocardial borders and blood pool of the left ventricle and left atrium were annotated. For PLAX views, the linear measurements were determined. For Doppler modalities, velocity trace and view-specific annotations were done. Based on the frame-level annotations, video-level volume curves were generated to identify end systole and end diastole (appendix p 11). End systole, end diastole, and peak positions were confirmed by automated analysis of the accompanying electrocardiogram, if present. We were then able to project frame-level annotations on videos in real time (video).

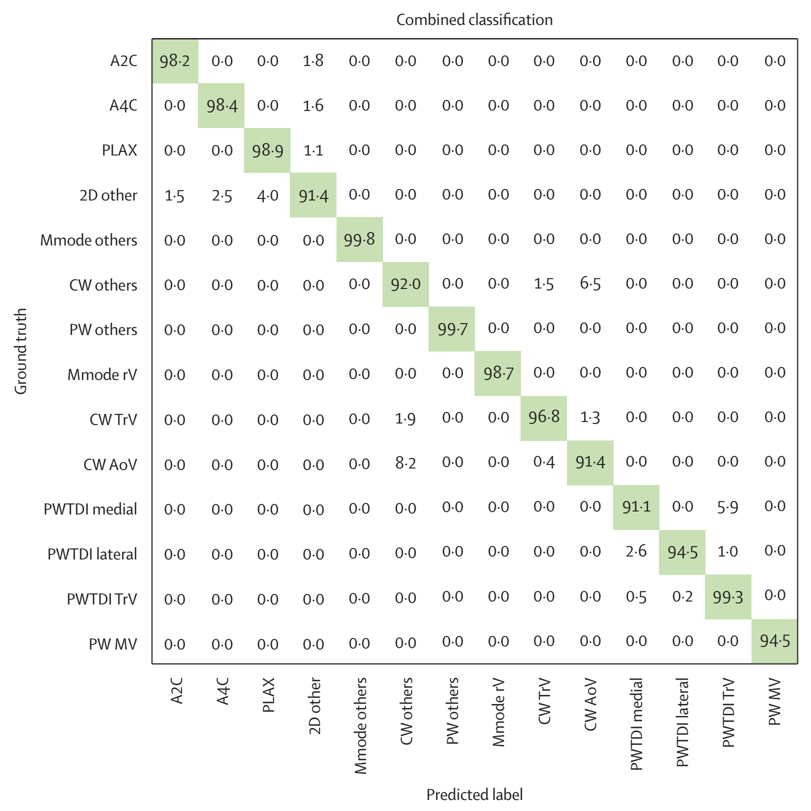
See **Online** for video

Filters and confidence score

Individual echocardiograms often contain multiple videos of the same views, and one video has numerous frames. Therefore, we used a confidence score to identify videos and frames of the best quality (view quality), and decision rules to identify automated measurements of the best quality (measurement quality), for the external validation sets. Videos of low quality were excluded from analyses. The view quality was based on the highest output probability of the view classifier CNN softmax layer, similar to previous attempts.[10] The measurement quality was based on several checks. These included the shape and placement of the annotation trace, congruency between systolic and diastolic phased with the electrocardiogram, and the automated measurement being within a physiological range.

Outlier identification and comparison with human variance

Three independent sonographers re-analysed the top 15 outliers (ie, echocardiograms with the greatest discordance between human and automated measurements) for each measurement in the Canadian dataset to identify causes of mismatch between automated and human measurements. The sonographers were masked to the original human measurement and the automatic measurement. Each sonographer remeasured volumes and visually assessed LVEF, E-wave, e' lateral and e' medial. After these assessments, the sonographer selected the closest value (ie, automated or original human measurement). The sonographers were also llowed to comment on the quality of the 2D video or



Ground truth \ Predicted label	A2C	A4C	PLAX	2D other	Mmode others	CW others	PW others	Mmode rV	CW TrV	CW AoV	PWTDI medial	PWTDI lateral	PWTDI TrV	PW MV
A2C	98·2	0·0	0·0	1·8	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
A4C	0·0	98·4	0·0	1·6	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
PLAX	0·0	0·0	98·9	1·1	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
2D other	1·5	2·5	4·0	91·4	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
Mmode others	0·0	0·0	0·0	0·0	99·8	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
CW others	0·0	0·0	0·0	0·0	0·0	92·0	0·0	0·0	1·5	6·5	0·0	0·0	0·0	0·0
PW others	0·0	0·0	0·0	0·0	0·0	0·0	99·7	0·0	0·0	0·0	0·0	0·0	0·0	0·0
Mmode rV	0·0	0·0	0·0	0·0	0·0	0·0	0·0	98·7	0·0	0·0	0·0	0·0	0·0	0·0
CW TrV	0·0	0·0	0·0	0·0	0·0	1·9	0·0	0·0	96·8	1·3	0·0	0·0	0·0	0·0
CW AoV	0·0	0·0	0·0	0·0	0·0	8·2	0·0	0·0	0·4	91·4	0·0	0·0	0·0	0·0
PWTDI medial	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	91·1	0·0	5·9	0·0
PWTDI lateral	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	2·6	94·5	1·0	0·0
PWTDI TrV	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·5	0·2	99·3	0·0
PW MV	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	94·5

***Figure 1:* Confusion matrix on the classification of 2D views and Doppler modalities for the Asia test set**
Values greater than 0 for non-shaded fields show misclassification. A2C=apical 2 chamber. A4C=apical 4 chamber. AoV=aortic outflow velocity. CW=continuous wave Doppler. MV=mitral valve. PLAX=parasternal long axis. PW=pulsed wave Doppler. PWTDI=pulsed wave tissue Doppler imaging. TrV=tricuspid regurgitation peak velocity. 2D=two dimensional.

Doppler modalities or whether the patient was in atrial fibrillation.

We also performed a prospective study to compare the variability of automated versus human measurements. Two independent expert sonographers remeasured up to 142 individual echocardiograms from ATTRaCT (n=115) and the Alberta HEART study (n=27). These images were not used in the training of the workflow. The sonographers each had more than 10 years of experience working in an echocardiography core laboratory. The two sonographers were presented with the same random selection of studies and were asked to measure left ventricular end systolic volume (LVESV), left ventricular end diastolic volume (LVEDV), LVEF, left atrial end systolic volume (LAESV), E-wave, e' lateral, and e' medial. They were masked to the original clinical measurements, each other's measurements, and the automated measurements. Variability between the two human measurements and the original clinical measurement (three human measurements in total) were compared with the variability between automated measurement and the three human measurements using the individual equivalence coefficient (IEC).

Statistical analysis

We calculated the mean absolute error (MAE), root mean square error, median absolute and relative (percentage) deviation, and the Pearson correlation coefficient, *r*, for automated versus manual measurements. For the internal test set we also present accuracy of measurements, which refers to the proportion of videos or images that were correctly classified in their respective categories for 2D videos or Doppler measurements, determined by the number of predictions divided by the total number with the same ground truth label. We performed post-hoc interaction analyses to test whether age, sex, or body-mass index (BMI) modified the association between automated and manual echocardiographic parameters for E/e' ratio, because of phenotypic heterogeneity of heart failure with preserved ejection fraction. The Dice coefficient of similarity was used to compare automated to manual annotations of cardiac chambers.[2] The absolute and percentage deviation were calculated for the 50th, 75th, and 95th percentiles of the automated measurements to compare automated and manual measurements. We evaluated the interchangeability of automated and human measurements in the prospective study using the reference-scaled IEC.[18] The IEC provides a metric to assess the variation between automated and human measurements compared with the variation between human measurements (ie, individual bio-equivalence). Thus, the expected value of IEC is 0 if automated and human measurements have identical within-patient variation, less than 0 if automated measurements have lower variation, and greater than 0 if automated measurements have higher variation compared with human measurements. The expected value of IEC is independent of population mean and intraclass correlation. We calculated the area under the receiver operating characteristic curve (AUC) for the ability of automated measurements to identify patients with systolic dysfunction (LVEF <40%) or having E/e' ratio of at least 13 or e' lateral of less than 10 cm/s,[2] on the basis of the original clinical (ground truth) measurements. In post-hoc analyses, we compared the performance of the workflow in patients with atrial fibrillation versus those without atrial fibrillation. The deep learning workflow was developed with Python (version 2.8); testing and validation of automated measurements was done with R (version 3.4.1).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Following the development and training of the automated workflow for echocardiographic analyses in the training dataset from Asia (ATTRaCT programme), we assessed the workflow performance in an internal test set and in separate, previously unseen, independent validation datasets.

In the test set from Asia, CNNs distinguished different views on 2D videos and Doppler modalities with an accuracy ranging from 91·1% for PWTDI medial (medial e') to 98·9% for PLAX (figure 1). CNNs were able to segment cardiac chambers with a mean Dice similarity coefficient (a measure of similarity of the annotations) ranging from 93·0% to 94·3% for both the left atrium and left ventricle (appendix p 5). The correlation between automated measurements with the manual measurements ranged from *r*=0·88 for E wave (MAE 7·4 cm/s), to a correlation of *r*=0·95 for LVESV (MAE 10·2 mL; table 1). For the most clinically relevant parameters, the correlation between automated and manual measurements was *r*=0·89 (MAE 5·5%) for LVEF, *r*=0·92 (MAE 0·7 cm/s) for e' lateral, and *r*=0·90 (MAE 1·7) for E/e' ratio (table 1, figure 2A–C). The AUC was 0·96 (95% CI 0·92–0·99) for identifying participants with systolic dysfunction (LVEF <40%), 0·95 (0·88–0·99) for an e' lateral wave velocity less than 10 cm/s, and 0·96 (0·92–0·99) for an E/e' ratio of 13 or higher (figure 3A–C). The association between ground truth E/e' ratio and automated E/e' measurements was not influenced by age, BMI, or sex in post-hoc interaction analysis ($p_{\text{interaction}}$ >0·10).

We performed external validation of the workflow in three datasets: a curated dataset from Canada (Alberta HEART Study), a real-world dataset from Taiwan (Mackay Memorial Hospital), and a reference dataset from the USA (EchoNet-Dynamic dataset). In the cohort from Canada, 0–2·0% of the 2D videos and Doppler modalities were of low view quality, and 1·3–10·9% were of low measurement quality (appendix p 6). Correlations between automated and manual measurements ranged from *r*=0·67 for e' medial (MAE 1·0 cm/s) to *r*=0·91 for LVESV (MAE 16·5 mL; table 1). The correlation between automated and manual measurements was *r*=0·75 (MAE 8·6%) for LVEF, *r*=0·78 (MAE 1·2 cm/s) for e' lateral, and *r*=0·75 (MAE 2·2) for E/e' ratio (table 1, figure 2A–C). The AUC was 0·91 (0·88–0·94) for identifying participants with LVEF less than 40%, 0·88 (0·84–0·92) for an e' lateral velocity less than 10 cm/s, and 0·91 (0·88–0·94) for an E/e' ratio of 13 or higher based on automated measurements (figure 3A–C).

In the dataset from Taiwan, 0–2·9% of 2D and Doppler modality images were of low view quality, and 1·3–28·1% were of low measurement quality (appendix p 8). Correlations between automated and manual measurements ranged from *r*=0·62 for LAESV (MAE 9·2 ml) to *r*=0·88 for e' lateral (MAE 1·6 cm/s; table 1). The correlation between automated and manual measurements was *r*=0·75 (MAE 10·2%) for LVEF, *r*=0·87 (MAE 1·6 cm/s) for e' lateral, and *r*=0·79 (MAE 1·8) for E/e' ratio (table 1, figure 2A–C). The AUC was 0·90 (0·89–0·90) for identifying participants with LVEF less than 40%, 0·94 (0·93–0·95) for an e' lateral velocity less than 10 cm/s, and 0·91 (0·89–0·93) for an E/e' ratio of 13 or higher (figure 3A–C).

The MAEs of measurements were higher in patients with atrial fibrillation than in patients without atrial fibrillation in the Canada and Taiwan cohorts (appendix pp 7, 9). However, *r* values were similar or higher in patients with atrial fibrillation for LVESV, LVEDV, LVEF, LAESV, and E/e' ratio in the Canadian cohort, and LVESV and LVEDV in the Taiwanese cohort.

We validated LVEF measurements in the US EchoNet-Dynamic dataset, which included 10 030 clinically

	Parameter				Absolute deviation *vs* manual measurement (% of manual)		
	Number of echo-cardiograms	*r*	Mean absolute error*	Root mean square error*	50th percentile	75th percentile	95th percentile
ATTRACT, Asia							
LVESV, mL	145	0·95	10·2	14·5	6·7 (9·3%)	15·5 (16·7%)	28·4 (29·1%)
LVEDV, mL	142	0·93	13·2	17·8	10·3 (7·9%)	17·5 (13·3%)	36·4 (22·0%)
LVEF, %	142	0·89	5·5	6·8	4·9 (12·9%)	8·2 (23·5%)	12·3 (49·3%)
LAESV, mL	162	0·93	5·1	7·3	3·5 (7·5%)	7·6 (13·9%)	13·4 (26·5%)
E wave, cm/s	258	0·88	7·4	12·6	3·8 (5·1%)	7·8 (10·3%)	30·4 (33·6%)
e' lateral, cm/s	235	0·92	0·7	1·1	0·4 (5·9%)	0·8 (13·3%)	2·2 (35·0%)
e' medial, cm/s	230	0·89	0·5	0·9	0·4 (6·5%)	0·5 (11·9%)	1·9 (42·2%)
E/e' ratio	157	0·90	1·7	2·7	1·2 (10·3%)	2·1 (16·5%)	4·4 (31·9%)
Alberta HEART study, Canada							
LVESV, mL	336	0·91	16·5	22·5	13·2 (27·4%)	21·5 (65·0%)	40·8 (138·0%)
LVEDV, mL	334	0·86	24·6	32·0	20·2 (20·9%)	31·4 (36·9%)	59·5 (79·5%)
LVEF, %	748	0·75	8·6	10·8	7·2 (13·5%)	12·0 (22·1%)	21·8 (40·3%)
LAESV, mL	714	0·88	10·8	15·5	7·9 (12·6%)	14·4 (21·4%)	29·6 (40·1%)
E wave, cm/s	420	0·81	11·3	17·8	7·8 (10·1%)	13·9 (17·6%)	34·6 (40·8%)
e' lateral, cm/s	400	0·78	1·2	1·9	0·7 (9·8%)	1·6 (20·4%)	3·9 (59·8%)
e' medial, cm/s	386	0·67	1·0	1·7	0·5 (8·3%)	1·1 (19·9%)	3·4 (49·2%)
E/e' ratio	598	0·75	2·2	3·8	1·4 (13·9%)	2·6 (23·1%)	6·7 (46·6%)
MacKay Memorial Hospital, Taiwan							
LVESV, mL	16 989	0·83	13·2	20·7	8·1 (20·0%)	16·0 (34·2%)	45·3 (66·5%)
LVEDV, mL	16 939	0·75	25·3	32·3	21·0 (20·3%)	35·6 (31·9%)	63·1 (52·7%)
LVEF, %	7724	0·75	10·2	12·6	8·9 (18·1%)	14·5 (29·5%)	24·3 (50·0%)
LAESV, mL	1892	0·62	9·2	11·8	7·7 (23·9%)	12·6 (40·5%)	22·7 (76·4%)
E wave, cm/s	18 659	0·71	11·6	18·9	6·7 (9·8%)	14·5 (20·3%)	40·4 (46·6%)
e' lateral, cm/s	6348	0·87	1·6	2·1	1·4 (16·7%)	2·1 (26·5%)	3·8 (51·2%)
e' medial, cm/s	4654	0·87	1·4	1·7	1·1 (15·7%)	1·7 (25·0%)	3·2 (50·0%)
E/e' ratio	2999	0·79	1·8	2·6	1·3 (16·7%)	2·4 (25·5%)	5·1 (43·0%)

Percentiles are for the absolute and relative devation. LAESV=left atrial end systolic volume. LVEDV=left ventricular end diastolic volume. LVEF=left ventricular ejection fraction. LVESV=left ventricular end systolic volume. *Units of measurement are listed in the left column.

Table 1: **Correlation coefficients for ground truth and automated measurements**

measured LVEF values. We identified 6476 A4C views for measurement of LVEF. Of these, 6306 (97·4%) had a high view quality, and 6286 (97·1%) had a high quality measurement, and thus 6286 views were assessed. The correlation between automated and manual measurements was *r*=0·76 (MAE 6·5%; appendix p 12). The AUC was 0·92 (0·91–0·94) for identifying participants with LVEF less than 40% (appendix p 12).

Among the top 15 outliers for each measurement in the Canadian dataset, the three independent sonographers preferred the automated measurements over the original human measurement for a mean of 84% LVEDVs, 74% LVESVs, 74% LAESVs, 63% e' medial velocities, 56% e' lateral velocities, 42% LVEFs, and 30% E-wave measurements (appendix p 13). When the original human measurement of LVEF was preferred over automated measurements by the clinical expert, video quality was often poor (six of 15 videos considered poor quality; manual measurement preferred in five of those six). When the human measurement of E-wave was preferred over automatic measurements by the clinical expert, patients were commonly in atrial fibrillation (five of 15 videos labelled with atrial fibrillation; manual measurement preferred in four of those five).

In the prospective validation of automated measurement versus expert human measurement of echocardiograms from the Asian (ATRaCT) and Canadian datasets, all IECs were less than 0 for all measurements, indicating that automated measurements were interchangeable with human measurements (table 2). The reference-scaled values of automated measurements were similar to the scaled values of human measurements (appendix p 14).

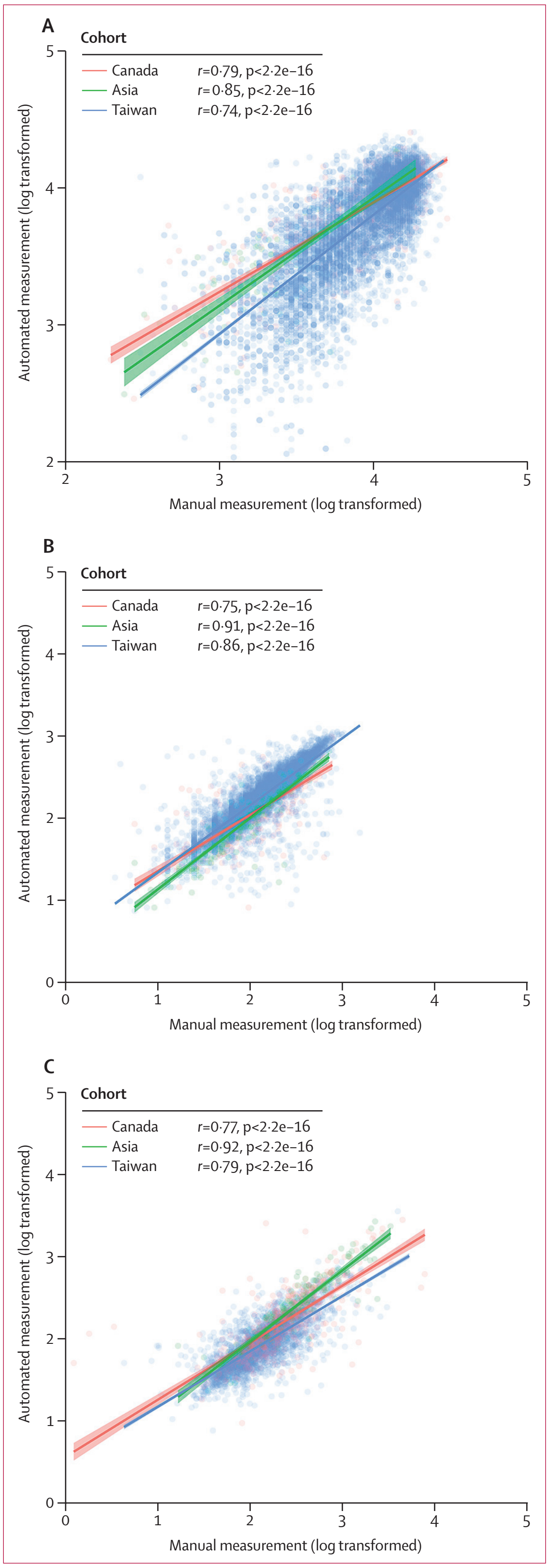


Figure 2: **Scatter plots with regression lines for left ventricular ejection fraction (A), e′ lateral (B), and E/e′ ratio (C)**
The *r* coefficients and p values are for the log transformed values; non-log transformed *r* values are shown in table 1.

Discussion

In this large-scale study with validation across four separate cohorts, we showed that a deep learning-based end-to-end workflow could automatically classify echocardiographic views and Doppler modalities and assess cardiac systolic and diastolic function parameters. The external validation of the workflow in distinct cohorts from different countries, health-care systems, and participants from Asia, Canada, and the USA showed the generalisability of automated measurements in men, women, and real-world clinical patients, with diverse ethnic backgrounds. We also showed that the variability of automatic measurements was lower than variability among manual measurements by expert sonographers.

Previous attempts that used deep learning to automate the annotation of echocardiograms focused on view classification,[8,9,19–23] assessment of systolic function,[11,13] or a limited number of measurements without Doppler modalities.[10] Notably, most studies had no external validation.[8,10,20–22,24] Recently, Ouyang and colleagues reported a deep learning algorithm for the automatic assessment of LVEF and validated this in an external dataset from a different health-care system.[11] However, this study only focused on evaluating LVEF, and whether the performance of this previous algorithm would be different in more diverse populations particularly in terms of ethnicity is unclear. Interpretation of cardiac systolic function is commonly done by health-care professionals with limited additional training, for example, in the emergency setting.[25] However, interpretation of other parameters, including cardiac Doppler measurements, requires considerable training and time investment, which might not always be readily available outside the cardiology department or in resource-limited settings. We used a combination of CNNs and ensemble models trained for their specialised views and tasks to annotate 2D images and Doppler modalities. CNNs have been used previously to classify Doppler images without providing annotations[26] or to annotate 2D videos with quantity volumes, left ventricular mass, and ejection fraction.[10,11] We extend on earlier work because our study combined automated view selection and automated annotation of 2D videos and Doppler modalities. Importantly, this pipeline can function as part of a learning health-care system by forming an integral part of clinical care while at the same time improving automated assessment by learning from the user.[27] Additionally, we extensively validated our workflow in datasets from different countries, health-care systems, and in both men and women from Asia and North America.

The variability in automated measurements was smaller than the variation of manual measurements by expert sonographers in our prospective study. This finding shows that deep learning algorithms can potentially substitute manual annotations of echocardiograms.[2,28] In the Alberta HEART study dataset from Canada and the real-world dataset from Mackay Memorial Hospital, Taiwan, the MAEs for LVEF were higher compared with those in the earlier attempt by Ouyang and colleagues.[11] This increase in MAE is likely to have been caused by differences in image quality and the proportion of possible inaccurate manual measurements. When expert sonographers examined the top 15 outliers for each measurement, they commonly preferred the automated measurements over the original human measurements. For those measurements which the original human measurement was preferred over automated measurement, image quality was often poor, or patients were often in atrial fibrillation. MAEs were higher in patients with atrial fibrillation than in those without, particularly for measurements influenced by the presence of atrial fibrillation, such as E-wave and LAESV. Poor view quality was a specific issue for the US EchoNet-dynamic dataset. Videos were compressed and of considerably worse quality than ordinary DICOM files. However, the MAE of automated LVEF measurements by our workflow in the EchoNet-Dynamic dataset was similar to the MAE of LVEF described by Ouyang and colleagues in external validation of their algorithm.[11] We used quality control criteria for the view classifier and measurements to select images of sufficient image quality. The advantage of this step is that it can improve the performance and usability of the workflow for end users. However, the limitation of this approach is that not all images will be annotated. The

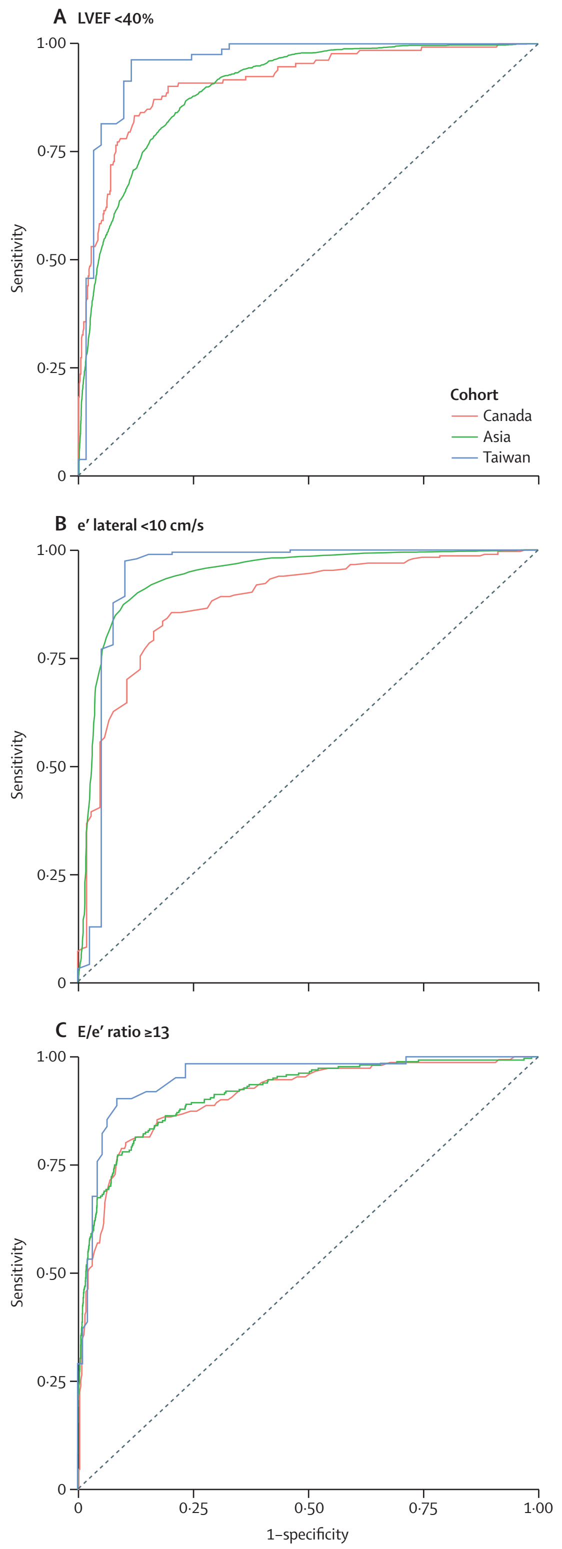


Figure 3: **AUC for identifying patients with clinical measurements of LVEF less than 40% (A), e′ lateral less than 10 cm/s (B), and E/e′ ratio at or higher than 13 (C) with automated measurements**
AUC=area under the receiver operating characteristic curve. LVEF=left ventricular ejection fraction.

	Number of echocardiograms with complete readings*	IEC
Left ventricular end systolic volume	45	–0·67
Left ventricular end diastolic volume	46	–0·33
Left ventricular ejection fraction	59	–0·54
Left atrial end systolic volume	53	–0·16
E wave	89	–0·6
e′ lateral	73	–0·28
e′ medial	76	–0·67
E/e′ ratio	32	–0·88

IEC=individual equivalence coefficient. *Three human measurements (two prospective measurements by sonographers and one original clinical measurement) and the automated measurement.

Table 2: **IECs for independent prospective validation between human expert measurements in triplicate *vs* automated measurements**

ability to verify the machine learning algorithms' output visually and manually could also help ease any concerns some might have with more opaque outcomes (commonly known as black box outcomes) used with alternative artificial intelligence-based approaches.

The present study is part of a broader paradigm shift in cardiovascular care. New deep learning approaches can augment or replace labour-intensive and repetitive tasks.[29] Deep learning algorithms can reduce interobserver and intraobserver variability and can be deployed at scale for the automated surveillance of echocardiographic databases. Combined with advances in the development of handheld echocardiographic devices, the provision of artificial intelligence software support in echocardiographic interpretation might increase access to cardiac imaging in settings in which clinical expertise is lacking, and resources are scarce.[30]

The present work's value is best understood against the background of its limitations. First, the presented workflow was trained on expert annotations by trained sonographers. Therefore, the present workflow can only analyse and annotate echocardiograms of sufficient quality. Training and validation were performed against expert human measurements, and we did not compare automated measurements to non-echocardiographic measurements, such as invasive haemodynamic measurements or magnetic resonance imaging. Therefore, our automated measurements reflect expert measurements for echocardiography. We validated our automated workflow against available measurements in cohorts from Canada and Taiwan, and the USA for LVEF. The availability of measurements might have introduced bias into the external validation. However, automated measurements had similar or better reproducibility and variability than manual measurements by expert sonographers in our prospective study, suggesting that possible bias did not severely affect our external validation. Although the number of echocardiograms of inadequate quality was low, further work is needed to identify how lower quality videos can be analysed, or how machine learning can be used upstream to guide the acquisition of good quality images.

We presented a fully automated deep learning-based workflow to automate the view classification, annotation, and interpretation of cardiac volumes, LVEF, and E/e' ratio. Our results are an important step forward and highlight the possibility of deep learning to provide a fully automated solution for interpreting echocardiograms, which can support clinicians and augment clinical care.

Contributors

All authors contributed equally to critical revisions of the Article and all authors accept final responsibility for the decision to submit for publication. JT, PJS, MBI, ZJ, WO, MJF, and HZ prepared the data and figures and performed statistical analyses. JT and PJS wrote the first draft of the Article with input from C-LH, JAE, CSPL, and SDS. The study was designed by all authors. Critical revisions of the Article were provided by C-LH, FE, RSMG, WH, L-HL, DS, PC, AMR, HKL, SDS, CSPL, and JAE. All authors had access to all the data, and JT, CSL and JAE have verified the data.

Declaration of interests

JT is supported by a National University of Singapore Startup Grant and has received personal grants and speaker fees from Olink Proteomics, Roche Diagnostics, Daiichi Sankyo, and Us2.ai. CSPL is supported by a Clinician Scientist Award from the National Medical Research Council of Singapore; has received research support from Boston Scientific, Bayer, Roche Diagnostics, AstraZeneca, Medtronic, and Vifor Pharma; has served as consultant or on the advisory board, steering committee, or executive committee for Boston Scientific, Bayer, Roche Diagnostics, AstraZeneca, Medtronic, Vifor Pharma, Novartis, Amgen, Merck, Janssen Research & Development, Menarini, Boehringer Ingelheim, Novo Nordisk, Abbott Diagnostics, Corvia, Stealth BioTherapeutics, JanaCare, Biofourmis, Darma, Applied Therapeutics, MyoKardia, Cytokinetics, WebMD Global, Radcliffe Group, and Corpus; and is co-founder and non-executive director of Us2.ai. MBI, MJF, and ZJ are employees of Us2.ai and hold equity in the company. PJS was an employee of Us2.ai and held equity in the company during the study. JT, PJS, WO, and CSPL own a patent entitled "Automatic clinical workflow that recognises and analyses 2D and doppler modality echocardiogram images for automated cardiac measurements and the diagnosis, prediction and prognosis of heart disease" related to the present work. JT, WO, and JAE are scientific advisors of Us2.ai and hold equity in the company. AMR is supported by a Senior Translational Research award from the National Medical Research Council of Singapore; holds the New Zealand Heart Foundation Chair of Cardiovascular Studies; has received research support from Boston Scientific, Bayer, AstraZeneca, Medtronic, Roche Diagnostics, Abbott Laboratories, Thermo Fisher Scientific, and Critical Diagnostics; and has consulted for Bayer, Novartis, Merck, AstraZeneca, and Roche Diagnostics. SDS reports grants from Actelion, Alnylam, Amgen, AstraZeneca, Bellerophon, Bayer, Bristol Myers Squibb, Celladon, Cytokinetics, Eidos, Gilead Sciences, GlaxoSmithKline, Ionis Pharmaceuticals, LoneStar Heart, Mesoblast, MyoKardia, NeuroTronik, NHLBI, Novartis, Novo Nordisk, Respicardia, Sanofi Pasteur, and Theracos; and personal fees from Abbott, Actelion, Akros, Alnylam Pharmaceuticals, Amgen, Arena, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol Myers Squibb, Cardior Pharmaceuticals, Cardurion Pharmaceuticals, Corvia Medical, Cytokinetics, Daiichi Sankyo, Gilead Sciences, GlaxoSmithKline, Ironwood Pharmaceuticals, Eli Lilly, Merck, MyoKardia, Novartis, Roche, Takeda, Theracos, Quantum Genetics, Cardurion Pharmaceuticals, AOBiome, Janssen, Cardiac Dimensions, Sanofi Pasteur, Tenaya, DiNAQOR, Tremeau Pharmaceuticals, CellProthera, and Moderna. All other authors declare no competing interests.

Data sharing

The external EchoNet-Dynamic dataset is publicly available. Access to the ATTRaCT dataset is open on reasonable request to the principal investigator of ATTRaCT, but restrictions apply to its availability (for reproducing the present work) and ethical approvals. Access to the Alberta HEART study and data from Mackay Memorial Hospital might be available after contacting their respective principal investigators and subject to ethical approvals for reproducing the current work.

Acknowledgments

We gratefully acknowledge Yoran M Hummel and James O Hare from Us2.ai for helpful discussions and logistical support, and Su Huang, Yi Xuan Chia, and Shoulie Xie from A*STAR for helpful discussions. The ATTRaCT programme was supported by research grants from the A*STAR Biomedical Research Council (grant numbers SPF2014/003, SPF2014/004, SPF2014/005) and A*STAR Exploit Technologies (grant number ETPL/17-GAP012-R20H). The Alberta Alberta HEART programme was supported by an Alberta Innovates–Health Solutions Interdisciplinary Team grant (grant number AHFMR ITG 200801018).

Editorial note: the *Lancet* Group takes a neutral position with respect to territorial claims in published tables and text and institutional affiliations.

References

1 Zannad F. Rising incidence of heart failure demands action. *Lancet* 2018; **391:** 518–19.

2 Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging* 2015; **16:** 233–70.

3 Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* 2016; **37:** 2129–200.

4 Yancy CW, Jessup M, Bozkurt B, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *Circulation* 2017; **136:** e137–61.

5 Nagueh SF, Smiseth OA, Appleton CP, et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging* 2016; **17:** 1321–60.

6 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

7 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 2199–210.

8 Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med* 2018; **1:** 6.

9 Khamis H, Zurakhov G, Azar V, Raz A, Friedman Z, Adam D. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Med Image Anal* 2017; **36:** 15–21.

10 Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* 2018; **138:** 1623–35.

11 Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020; **580:** 252–56.

12 Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *npj Digit Med* 2020; **3:** 1–10.

13 Asch FM, Poilvert N, Abraham T, et al. Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert. *Circ Cardiovasc Imaging* 2019; **12:** e009303.

14 Nagueh SF, Abraham TP, Aurigemma GP, et al. Interobserver variability in applying American Society of Echocardiography/European Association of Cardiovascular Imaging 2016 guidelines for estimation of left ventricular filling pressure. *Circ Cardiovasc Imaging* 2019; **12:** e008122.

15 Paulus WJ, Tschöpe C, Sanderson JE, et al. How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. *Eur Heart J* 2007; **28:** 2539–50.

16 Ezekowitz JA, Becher H, Belenkie I, et al. The Alberta Heart Failure Etiology and Analysis Research Team (HEART) study. *BMC Cardiovasc Disord* 2014; **14:** 91.

17 Guo X, Liu X, Zhu E, Yin J. Deep clustering with convolutional autoencoders. In: Liu D, Xie S, Li Y, Zhao D, El-Alfy ES (eds). Neural Information Processing. Springer, 2017: 373–82.

18 Barnhart HX, Kosinski AS, Haber MJ. Assessing individual agreement. *J Biopharm Stat* 2007; **17:** 697–719.

19 Balaji GN, Subashini TS, Chidambaram N. Automatic classification of cardiac views in echocardiogram using histogram and statistical features. *Procedia Comput Sci* 2015; **46:** 1569–76.

20 Liao Z, Jafari MH, Girgis H, et al. Echocardiography view classification using quality transfer star generative adversarial networks. In: Shen D, Liu T, Peters TM, et al (eds). Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Springer, 2019: 687–95.

21 Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules* 2020; **10:** E665.

22 Park JH, Zhou SK, Simopoulos C, Otsuki J, Comaniciu D. Automatic cardiac view classification of echocardiogram. IEEE 11th International Conference on Computer Vision; Oct 14–21, 2007.

23 Zhang Z, Seibold H, Vettore MV, Song W-J, François V. Subgroup identification in clinical trials: an overview of available methods and their implementations with R. *Ann Transl Med* 2018; **6:** 122–122.

24 Lang RM, Addetia K, Miyoshi T, et al. Use of machine learning to improve echocardiographic image interpretation workflow: a disruptive paradigm change? *J Am Soc Echocardiogr* 2021; **34:** 443–45.

25 Randazzo MR, Snoey ER, Levitt MA, Binder K. Accuracy of emergency physician assessment of left ventricular ejection fraction and central venous pressure using echocardiography. *Acad Emerg Med* 2003; **10:** 973–77.

26 Gilbert A, Holden M, Eikvil L, et al. Doppler spectrum classification with CNNs via heatmap location encoding and a multi-head output layer. *arXiv* 2019; published online Nov 6. http://arxiv.org/abs/1911.02407 (preprint).

27 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

28 Thorstensen A, Dalen H, Amundsen BH, Aase SA, Stoylen A. Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study. *Eur J Echocardiogr* 2010; **11:** 149–56.

29 Zhu H, Cheng C, Yin H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit Health* 2020; **2:** e348–57.

30 Chamsi-Pasha MA, Sengupta PP, Zoghbi WA. Handheld echocardiography: current state and future perspectives. *Circulation* 2017; **136:** 2178–88.