2-dimensional echocardiographic Global Longitudinal Strain with Artificial Intelligence using Open Data from a UK-wide collaborative

Stowell, Catherine C., Howard, James P., Ng, Tiffany, Cole, Graham D., Bhattacharyya, Sanjeev, Sehmi, Jobanpreet, Alzetani, Maysaa, Demetrescu, Camelia D., Hartley, Adam, Singh, Amar, Ghosh, Arjun, Vimalesvaran, Kavitha, Mangion, Kenneth, Rajani, Ronak, Rana, Bushra S., Zolgharni, Massoud ORCID logoORCID: https://orcid.org/0000-0003-0904-2904, Francis, Darrel P. and Shun-Shin, Matthew J. ORCID logoORCID: https://orcid.org/0000-0002-1179-0867 (2024) 2-dimensional echocardiographic Global Longitudinal Strain with Artificial Intelligence using Open Data from a UK-wide collaborative. JACC: Cardiovascular Imaging, 17 (8). pp. 865-876. ISSN 1936-878X

https://doi.org/10.1016/j.jcmg.2024.04.017

JACC: CARDIOVASCULAR IMAGING VOL. 17, NO. 8, 2024

ORIGINAL RESEARCH

2-Dimensional Echocardiographic Global Longitudinal Strain With Artificial Intelligence Using Open Data From a UK-Wide Collaborative

Catherine C. Stowell, MSc,[a,*] James P. Howard, MA, MRCP, PhD,[a,*] Tiffany Ng, MSc, MRCP,[a] Graham D. Cole, MA, MRCP, PhD,[b] Sanjeev Bhattacharyya, MD,[c] Jobanpreet Sehmi, BSc, PhD,[d] Maysaa Alzetani, MRCP, MSc (Cardio), MSc (Med), MRCP(UK),[e] Camelia D. Demetrescu, BSc, HCPC,[f] Adam Hartley, MBBS, BSc, MRCP,[a] Amar Singh, MBBCHir,[g] Arjun Ghosh, MBBS, MSc, PhD,[h] Kavitha Vimalesvaran, MBBS, MSc, MRCP,[a] Kenneth Mangion, MD, MRCP, PhD,[i] Ronak Rajani, BM, MD,[j] Bushra S. Rana, MBBS,[k] Massoud Zolgharni, BSc, MPhil, PhD,[l] Darrel P. Francis, MA, MD,[a] Matthew J. Shun-Shin, BM, BCh, PhD[a]

ABSTRACT

BACKGROUND Global longitudinal strain (GLS) is reported to be more reproducible and prognostic than ejection fraction. Automated, transparent methods may increase trust and uptake.

OBJECTIVES The authors developed open machine-learning–based GLS methodology and validate it using multiexpert consensus from the Unity UK Echocardiography AI Collaborative.

METHODS We trained a multi-image neural network (Unity-GLS) to identify annulus, apex, and endocardial curve on 6,819 apical 4-, 2-, and 3-chamber images. The external validation dataset comprised those 3 views from 100 echocardiograms. End-systolic and -diastolic frames were each labelled by 11 experts to form consensus tracings and points. They also ordered the echocardiograms by visual grading of longitudinal function. One expert calculated global strain using 2 proprietary packages.

RESULTS The median GLS, averaged across the 11 individual experts, was −16.1 (IQR: −19.3 to −12.5). Using each case's expert consensus measurement as the reference standard, individual expert measurements had a median absolute error of 2.00 GLS units. In comparison, the errors of the machine methods were: Unity-GLS 1.3, proprietary A 2.5, proprietary B 2.2. The correlations with the expert consensus values were for individual experts 0.85, Unity-GLS 0.91, proprietary A 0.73, proprietary B 0.79. Using the multiexpert visual ranking as the reference, individual expert strain measurements found a median rank correlation of 0.72, Unity-GLS 0.77, proprietary A 0.70, and proprietary B 0.74.

CONCLUSIONS Our open-source approach to calculating GLS agrees with experts' consensus as strongly as the individual expert measurements and proprietary machine solutions. The training data, code, and trained networks are freely available online. (JACC Cardiovasc Imaging 2024;17:865-876) © 2024 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

From the [a]National Heart & Lung Institute, Imperial College, London, United Kingdom; [b]Department of Cardiology, Charing Cross Hospital, London, United Kingdom; [c]Department of Cardiology, St Bartholomew's Hospital, London, United Kingdom; [d]Department of Cardiology, West Hertfordshire Hospitals NHS Trust, Watford, United Kingdom; [e]Department of Cardiology, Luton & Dunstable University Hospital, Bedfordshire, United Kingdom; [f]Department of Cardiology, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom; [g]Department of Cardiology, Lewisham & Greenwich NHS Trust, London, United Kingdom; [h]Barts Heart Centre and Hatter Cardiovascular Institute, University College London Hospital, London, United

ISSN 1936-878X https://doi.org/10.1016/j.jcmg.2024.04.017

JACC: CARDIOVASCULAR IMAGING, VOL. 17, NO. 8, 2024
AUGUST 2024:865–876

ABBREVIATIONS AND ACRONYMS

AI = artificial intelligence

GLS = global longitudinal strain

LV = left ventricular

Global longitudinal strain (GLS) of the left ventricle is known to detect subtle reductions of left ventricular (LV) function across a spectrum of diseases, such as valvular heart disease,[1] chemotoxicity,[2] and ischaemia.[3,4] There are many methods of measuring this strain, with the commonest being speckle tracking using a proprietary system provided by the ultrasound vendor.[5]

Clinicians want measurement protocols to be designed and debated openly by their representative expert bodies. This allows protocols to be refined over the years, based on peer-reviewed published research. Clinicians train and are tested to match this expert consensus methodology as a part of their certification and revalidation processes. Proprietary systems do not have the advantages of open development by practicing clinicians, and of explicit methodology, that can be found to be incorrect and then improved. One reason for the lack of clinical uptake of GLS[6] is the absence of this same clinical grounding that gives clinicians confidence in their other measures.

We set out to solve these problems so that GLS can deliver in clinical practice its potential to increase the accuracy, reproducibility, and time efficiency of the assessment of LV systolic function. We applied 3 principles.

First, we assembled an expert reference group of practicing clinical echocardiographers, who provided not only the gold standard (by their consensus), but also an indication of the degree of divergence from expert consensus that is clinically acceptable for an expert.

Second, we designed an open GLS measurement system that does not rely on tracking individual speckles. The openness of the system permits clinical researchers to find and improve weaknesses. The nonreliance on speckle tracking allows it to operate successfully even when image quality and framerate does not permit following of a speckle from frame to frame.

Third, we used an image dataset from consecutive echocardiograms from our clinical service, not selected for image quality.

In this article, we present an approach to using supervised machine learning to deliver: 1) a 4,948-case, publicly accessible and representative video dataset for system development; 2) expert labelling of the endocardial border for 6,819 frames (3,218 apical 4-chamber, 1,164 apical 2-chamber, 2,437 apical 3-chamber) drawn from these 4,948 videos; 3) a separate 100-case video dataset for external validation, with each case having, for an end systolic and end-diastolic frame, 11 independent, mutually blinded expert tracings of the endocardial borders; 4) conventional measurements of GLS on the external validation dataset using 2 commercial proprietary software packages; 5) the trained neural networks and associated source code for this task; and 6) the results of the performance of the machine learning methods against the expert consensus gold standard, but with the individual expert opinions available to provide a context of how much variation from the consensus is acceptable for an expert.

METHODS

DATASETS. We created 2 datasets, 1 for model development and 1 for external validation. The model development dataset was used for training and internal validation. It consisted of 4,948 video loops covering the apical 4-chamber, apical 2-chamber, and apical 3-chamber views, from a range of manufacturers and machines.

The external validation dataset consisted of the apical 4-, 2-, and 3-chamber view of 100 patients acquired clinically over 3 consecutive working days outside of the time interval of the model development dataset (**Table 1**).

LABELLING KEY POINTS AND ENDOCARDIAL BORDER ON IMAGES. Each of the 6,819 model development images was labelled once. This was done by a pool of 36 experts, using the Unity UK Online Interface, which we have reported previously.[7] This web-based, interactive, real-time platform (**Figure 1**) allows experts to label the key points (septal and lateral mitral hinges, and endocardial apex) and the endocardial border. Details are given in

Kingdom; [i]School of Cardiovascular and Metabolic Health, University of Glasgow, Glasgow, United Kingdom; [j]Cardiovascular Directorate, St. Thomas' Hospital, King's College, London, United Kingdom; [k]Department of Cardiology, Hammersmith Hospital, London, United Kingdom; and the [l]School of Computing and Engineering, University of West London, London, United Kingdom. *Drs Stowell and Howard contributed equally to this paper.

Manuscript received August 28, 2023; revised manuscript received April 19, 2024, accepted April 25, 2024.

TABLE 1 Image Sources for Datasets

	Dataset 1: Training, Tuning, and Internal Validation Test Set	Dataset 2: External Validation Test Set
Dataset size	3,218 apical 4-chamber images from 2,162 video loops, 1,164 apical 2-chamber images from 387 video loops, and 2,437 apical 3-chamber images from 2,399 video loops were annotated by experts.	100 apical 4-chamber, 100 apical 2-chamber, 100 apical 3-chamber videos of which 11 experts labelled the 600 end-systolic and end-diastolic frames. The LV longitudinal function was also visually ranked.
Dataset source	Random selection of images from a 2-year period from 7 laboratories during 2015 and 2016	Sequential echocardiograms conducted over 3 days in 2019
Sex	Male: 401 (32.8%); female: 753 (61.5%); unspecified: 70 (5.7%)	Male: 47 (47%); female: 53 (53%)
Age, y	Median: 67 (IQR: 48-78)	Median: 60 (IQR: 48.5-73.0)
Year collected	2015 and 2016	2019
Manufacturer and model		
Philips	iE33: 297 Affinity 70C: 243 Epic 7C: 158 Affinity 50G: 32 CX50: 13	Single manufacturer compatible with proprietary software A and B
GE	Vivid i: 317 Vivid q: 145 Vivid S70: 12 Vivid S6: 4 Vivid E9: 2 Vivid 7: 1	

LV= left ventricular.

the Supplemental Methods, and a demonstration is available at the project website.[8]

For the external validation dataset, in contrast, a smaller group of 11 validator experts each labelled all 100 end-systolic and 100 end-diastolic images for each of the 3 apical views, mutually blinded to the labels of the others.

DEVELOPING AND TRAINING THE NEURAL NETWORK. The single-frame neural network was developed using the 6,819 model-development images. Approximately 70% of the images were used for training, 15% for tuning, and 15% for internal validation.

Neural networks can process images from videos through 2 broad approaches: examining only the frame of interest or having additional access to a few frames before and after.

The Unity-GLS is based on the convolutional neural network architecture, Higher HRnet-W32, which is designed to preserve fine spatial detail, having first been introduced for estimating poses from photographs of people. As an input it received the frame of interest plus 6 additional frames: at frame numbers −9, −3, −1, +1, +3, and +9. Where additional frames were missing (eg, at the beginnings and ends of videos), blank frames were substituted. A single network was trained for all 3 views: apical 4-, 3-, and 2-chamber.

All training images, labels, and associated code are available at the project website.[8]

VALIDATING THE NEURAL NETWORK. Both the experts and the neural networks drew the LV endocardial contour. For each video, the end-systolic and end-diastolic frames were drawn by a single expert in advance. The endocardial curve length of the left ventricle was taken from the septal mitral annular hinge point, via the LV endocardial apex, to the lateral mitral annular hinge point, following the endocardial contour. GLS was then calculated as the change in endocardial curve length from diastole to systole, expressed as a percentage of the end-diastolic curve length.

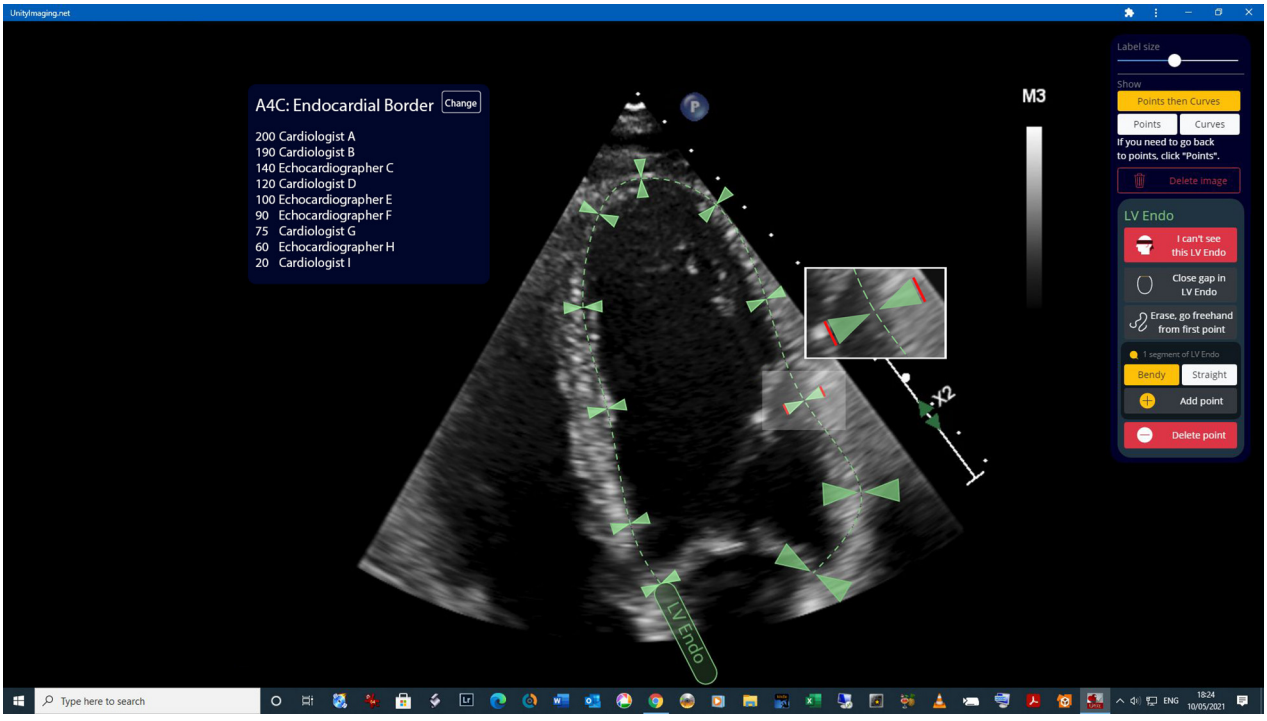
We calculated the GLS separately for each view and defined the GLS as the mean of the 3. For the expert consensus, we averaged 11 expert GLSs.

Strain analysis of the 100 patient studies was carried out by a single expert (CS), trained by the manufacturer representative. This was done separately using 2 proprietary software packages for GLS (Supplemental Methods).

Additional external validation was performed using visual longitudinal function. For this, the 11

JACC: CARDIOVASCULAR IMAGING, VOL. 17, NO. 8, 2024
AUGUST 2024:865–876

FIGURE 1 The Unity Web-Based Interface



The Unity online platform allows experts across the UK to easily and rapidly annotate medical images (identities on the leader board have been removed for publication). Experts use the platform to label key points and curves such as the endocardial border shown above. Details are given in the Supplemental Methods, and a demonstration is available online https://data.unityimaging.net.

experts worked as a team on each of the 3 sets of 100 external validation video loops, 1 set for each view. They ranked the videos of each view into order of subjective longitudinal function. The global rank was calculated as the average of the rank of the 3 views (Supplemental Methods).

STATISTICAL ANALYSIS. The primary statistical analyses were of single interpretations (from an expert, from a proprietary algorithm, or from each of the trained networks) against the consensus of experts. The consensus of experts was constructed from the median of the independent assessments made by the 11 experts of each of the 100 patient studies of the external validation dataset.

We define the error as the absolute difference between a single measurement of strain (either by an individual expert or a machine algorithm) and the value derived from the 11 experts. We present the pooled error of all 11 experts and the individual errors of the 4 machine algorithms. We also present the 50th (median), 75th, 80th, and 90th centiles of the error distribution. A study-wise comparison of median absolute errors was performed with the Mann-Whitney U test. For key point localization, the error is defined as the Euclidean distance between the individual's positioning and the consensus. For familiarity among clinicians, we also perform the equivalent of the Bland-Altman analysis, that is, separated the error into a bias component (the mean of the signed differences) and the residual noise component (the SD of the signed differences).

We present scatter plots of the assessments by each individual (expert or machine) against the expert consensus as a reference, and Spearman's correlation coefficient. Comparisons of dependent correlations were performed using the "cocor" statistical package[9] and Hotelling's T^2.[10] The intraclass correlation between the 11 validator experts was calculating using the patient and residual variance derived from a model with study as a random-effect. Analyses were performed using the statistical language and environment R.[11]

DATA AVAILABILITY. The complete training dataset consisting of images and annotations, as well as the code for the neural networks and trained models, are available online[8] and with additional information and

the PRIME (Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation) checklist[12] in the Supplemental Methods.

ETHICAL APPROVAL. Ethical approval was obtained from the South Central-Oxford C Research Ethics Committee (20/SC/0386).

RESULTS

DATASETS. We created 2 separate, nonoverlapping, expert-annotated datasets. The first was for the development process (primarily training the neural network). The second was for the external validation process (**Table 1**). Each image in the development dataset was labelled by a member of the pool of 36 experts. The external validation dataset comprised 100 consecutive studies conducted over 3 working days in 2019, several years away from the development dataset.

Each frame was then labelled by each of the 11 validator experts producing 6,600 potentially labelled frames. Overall, on 24 occasions the expert felt the image was of insufficient quality to label, so there were 6,576 labelled frames (99.6%). No image was refused by all experts.

Videos from 1 patient were unable to be analyzed using the proprietary methods. To present comprehensive measurements on a consistent patient group, we eliminated the videos of that patient and so the external validation results are based on 99 patients.

EXTERNAL VALIDATION: KEY POINT LOCATION AND ENDOCARDIAL BORDER LENGTH. For the external validation dataset, the consensus (median) of 11 experts provided the reference standard. Against this, we tested the individual experts and each of the 4 machine methods. We assessed the performance of each step of the process in calculating the GLS.

Identification of the apex and mitral hinge points is the first step. For the individual experts (using the consensus of experts as the reference standard), for the apical 4-chamber view, the median absolute error of positioning was 0.41 cm for the apex, 0.34 cm for the lateral hinge, and 0.31 cm for the septal hinge. For Unity-GLS they were 0.25 cm, 0.30 cm, and 0.32 cm, respectively. Full data for all 3 views are shown in Supplemental Tables 1 and 2.

The length of the endocardial curve through those points is the second step. For the individual experts, the median absolute was error was 0.76 cm in the apical 4-chamber view, 1.05 cm in the 2-chamber view, and 0.74 cm in the 3-chamber view. For Unity-GLS they were 0.47 cm, 1.80 cm, and 1.78 cm, respectively. Full data for all 3 views are shown in Supplemental Table 3.

EXTERNAL VALIDATION: GLS. For each of the study videos, the expert consensus manual GLS was defined as the average of the strain values calculated from each of the individual experts' labelling of the curve and points of the end-diastolic and end-systolic images.

The median expert consensus global strain of the validation dataset was −16.1% (IQR: −19.3% to −12.5%). The values of the 3 individual views and of the 3 machine methods are shown Supplemental Figure 1 and Supplemental Table 4.

Each individual expert's manually measured strain correlated with the expert consensus GLS, with a median correlation coefficient of 0.85 (IQR: 0.80-0.88) (**Figure 2**).

Each of the 3 machine methods of measuring strain correlated with the expert consensus strain: correlations 0.73 for proprietary A, 0.79 for proprietary B, and 0.91 for Unity-GLS (**Figure 2**). Unity-GLS had a higher correlation than all other methods ($P < 0.001$ for each comparison against Unity-GLS). The correlation between the 2 proprietary methods was 0.77.

The individual expert global strain measurements had a median absolute error of 2.0 GLS percentage units from the expert consensus global strain. The intraclass correlation within the 11 experts was 0.65. Of the machine methods, Unity-GLS performed best (median absolute error: 1.3), with the proprietary methods showing larger errors: proprietary A 2.5 ($P < 0.0001$) and proprietary B 2.2 ($P = 0.0001$).

Unity-GLS performed better ($P < 0.0001$) than the individual expert measurements, even though the reference standard was the consensus derived from those same expert measurements.

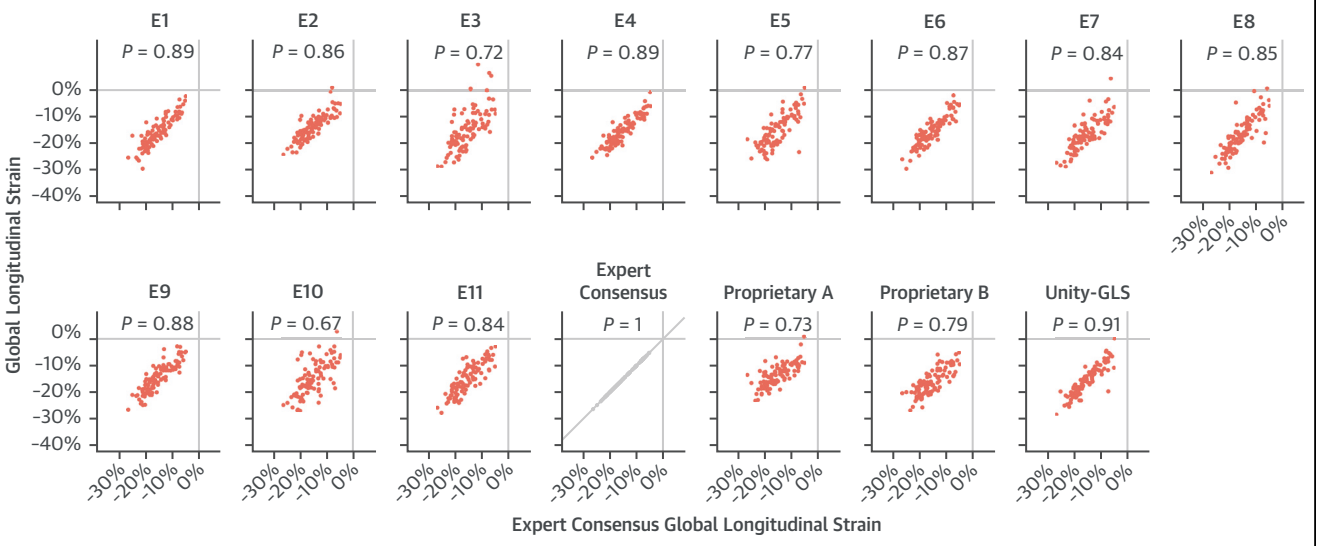
A Bland and Altman analysis and quantiles of absolute errors is shown in **Table 2**.

SEPARATE VALIDATION USING EYEBALL RANKING OF LONGITUDINAL FUNCTION. Strain measurements are a formalization of the observation from experts that different hearts show different amounts of longitudinal function. To test whether the machine methods had not drifted too far from this conceptual basis, the experts ranked the 100 cases for each view into the order of perceived longitudinal function, by sorting the individual views using the online system as described in the Supplemental Methods.

The individual expert GLS measurements showed a correlation with the consensus visual longitudinal

JACC: CARDIOVASCULAR IMAGING, VOL. 17, NO. 8, 2024
AUGUST 2024:865-876

FIGURE 2 Strength of Relationship Between Individual Experts or Individual Machine Methods, and the Consensus of Experts



(E1 to E11) Relationships between individual expert assessments of GLS (y-axis) and the consensus of experts (x-axis). Performance of the 3 machine methods (proprietary A, proprietary B, and Unity-GLS) are also shown on the y-axis. The Spearman rank correlations range from 0.67 to 0.89 for the individual experts and 0.73 to 0.91 for the machine methods. All $P < 0.001$. GLS = global longitudinal strain.

function rating of 0.72 (IQR: 0.68-0.75). The machine strains showed correlations of 0.77 for Unity-GLS, 0.70 for proprietary A, and 0.74 for proprietary B. These correlations were not significantly different from each other (Figure 3).

ORIGIN OF DISAGREEMENTS IN GLS MEASUREMENTS. In Supplemental Figure 2, we present all 594 frames from the 99 included patients of the validation dataset with the tracings from the 11 individual experts and from the Unity-GLS neural network (1 patient was excluded owing to incompatibility with the proprietary strain software).

A summary of representation of the best, worst, and average cases is given in Figure 4. There are 3 cases with best performance by Unity-GLS (top), median performance (middle), and worst (bottom) performances. Notably, the images in which Unity-GLS performed poorly were also images in which the experts disagreed with each other.

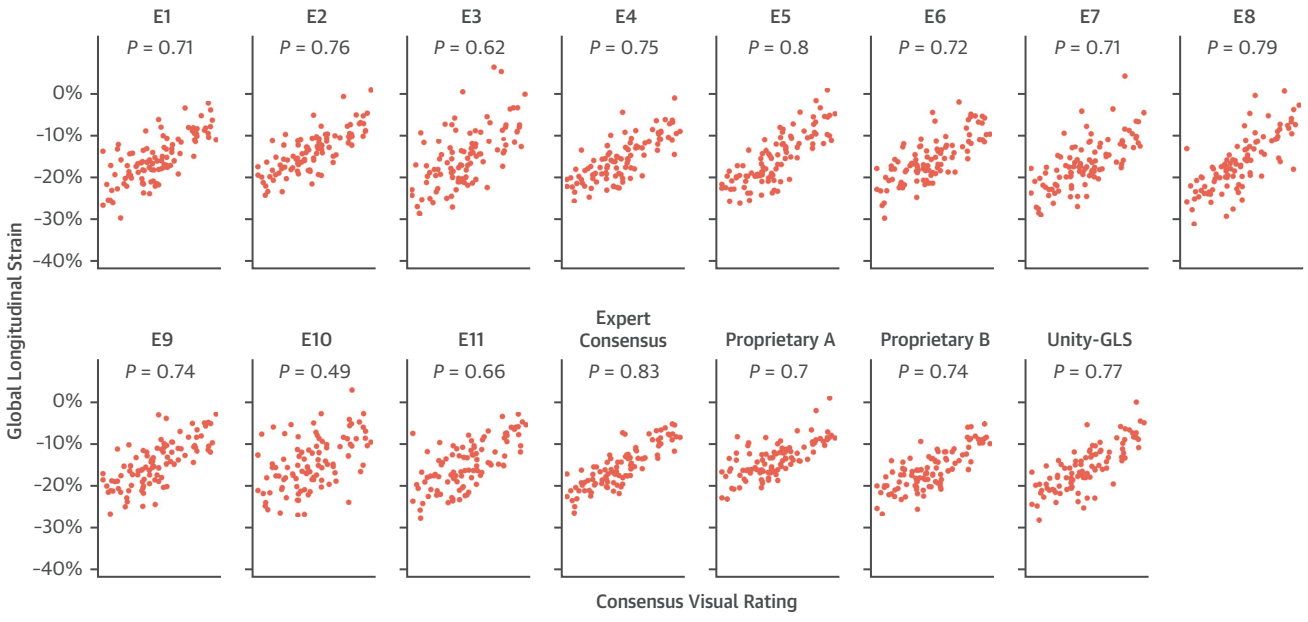
TABLE 2 Performance of the Individual Experts, Proprietary Methods, and Unity-GLS as Compared With the 11-Expert Consensus for Measuring LV GLS in the External Validation Dataset

	Bland-Altman Analysis		Quantile of Absolute Error From Expert Consensus			
Strain (Absolute %)	Mean Difference (95% CI) *P* value	SD of Difference	50% (Median)	75%	80%	90%
Individual experts	−0.15 (−0.35 to 0.05) $P = 0.15$	3.49	2.0	3.49	3.94	5.21
Proprietary A	−1.86 (−2.54 to −1.18) $P < 0.0001$	3.40	2.53	4.00	4.40	6.36
Proprietary B	0.54 (−0.07 to 1.17) $P = 0.09$	3.12	2.16	3.76	4.09	5.19
Unity-GLS	0.19 (−0.28 to 0.67) $P = 0.19$	2.39	1.33	2.23	2.48	3.45

Performance of the 3 machine methods and individual experts against the 11-expert consensus. For the Bland-Altman analysis of individual experts, every value by every individual was treated as a separate data point.

GLS = global longitudinal strain.

FIGURE 3 How Individual GLS Assessments Relate to Expert Consensus of a Visual Rating of Longitudinal Function



The pool of the 11 validator experts jointly ranked the videos by visually estimated left ventricular longitudinal function. Each panel shows how a single measurement of GLS (y-axis) correlates with that consensus visual rating (x-axis). The individual expert GLS assessments (E1 to E11) showed correlations with the consensus visual rating ranging from 0.49 to 0.80. The machine methods (proprietary A, proprietary B, and Unity-GLS) showed correlations from 0.70 to 0.77. The strongest relationship was between the expert consensus GLS and the expert consensus visual rating (0.83). All $P < 0.001$. Abbreviation as in Figure 2.

DISCUSSION

Clinical experts have been reticent to accept artificial intelligence (AI) in their workflows. One reason is that, unlike fellow clinical experts, neural networks do not have a clinically respected standard of training or ongoing open quality control. The Unity Collaborative was set up to tackle this in the UK by establishing a consortium of recognized experts to work together both to teach the neural network and to set the standards by which it is judged. A previous paper from the Unity Imaging Collaborative applied this process to automating the key clinically relevant measurements in the parasternal long axis view.[7] The present study applies the process to global LV strain.

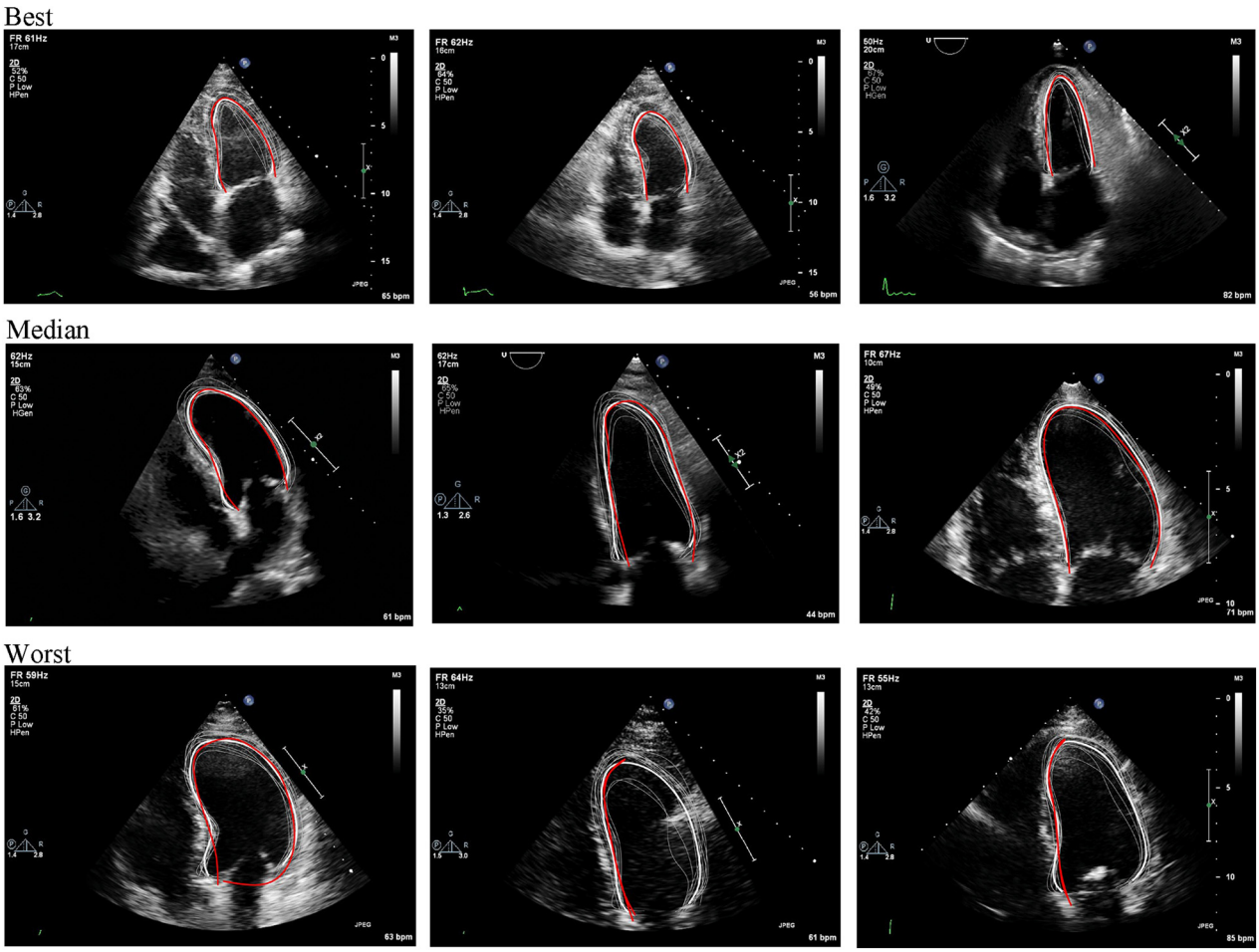
A key principle of our process for the separate external validation dataset is that the reference standard against which the machine methods were tested was not the opinion of a single expert, but the consensus across 11 experts who evaluated each image blinded to the evaluations of others (Central Illustration).

Having this multiexpert consensus allowed us to test the performance of all the machine methods against the performance of individual experts. This strategy showed the machine methods performed adequately in this task.

DIRECT MEASUREMENTS OF STRAIN WITHOUT SPECKLE TRACKING. Current technology for strain is reported by manufacturers to operate by tracking the motion of image speckles frame by frame, and then integrating the resulting velocities to derive the change in length. However, clinicians have not trusted this process enough to adopt it universally. For example, sometimes the fiducial points marked by the proprietary algorithms are seen to move in one direction, whereas the speckles seem to the naked eye to be moving in the opposite direction.

Because strain is ultimately a shortening of the LV border length,[13] it can in principle be measured directly from the end-systolic and end-diastolic frames without tracking speckles over the intervening frames. Direct measurements of strain have the advantage of being verifiable by human experts. It also allows for the potential of like-for-like comparisons. Clinicians can focus on the important task of identifying which cases it finds difficult and correcting its behavior, just as they do with human trainees.

FIGURE 4 9 Representative Apical 4-Chamber Cases Showing the Individual Experts, Expert Consensus, and the Output of Unity-GLS Tracing the LV Endocardium

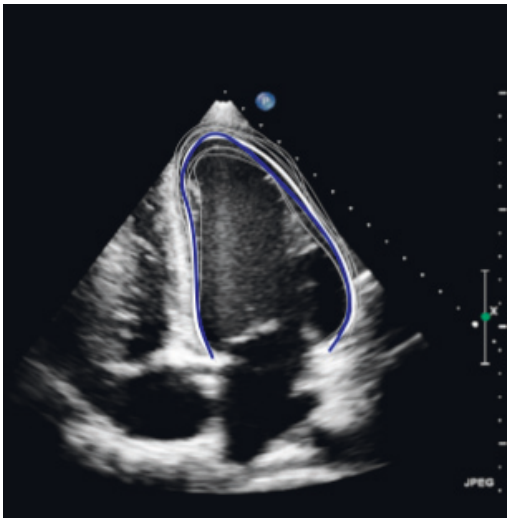


The 3 cases with the best (top), median (middle), and worst (bottom) agreement between the expert consensus curve (thick white line) and Unity-GLS (red line) output are shown. As a comparison, the individual experts' labels are shown (thin white lines). Images in which Unity-GLS performed poorly were also images in which the experts disagreed with each other. LV = left ventricular; other abbreviation as in Figure 2.

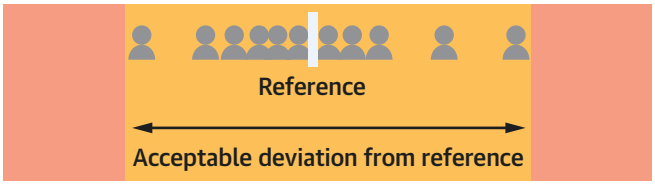
VALUE CONTRIBUTED BY OPEN RESEARCH. Many recent advances in automated echocardiogram interpretation have come from open research. Echocardiographic images labelled by experts, specifically for training neural networks, have been in short supply. In 2017, Smistad et al[14] introduced a method of to overcome this data shortage by pretraining the neural network on the output of an existing, non-neural network model, before refining it with a small amount of expert-labelled data. In 2019, Leclerc et al[15] applied a U-net to CAMUS, the then largest publicly available dataset, which contains 500 cases. The following year, Wei et al[16] introduced simultaneous learning of both segmentation masks and motion fields so that each process strengthened the performance of the other. Recently, Painchaud et al[17] innovated the use of variational autoencoders to automatically detect and improve the boundaries that the original neural network had drawn implausibly.

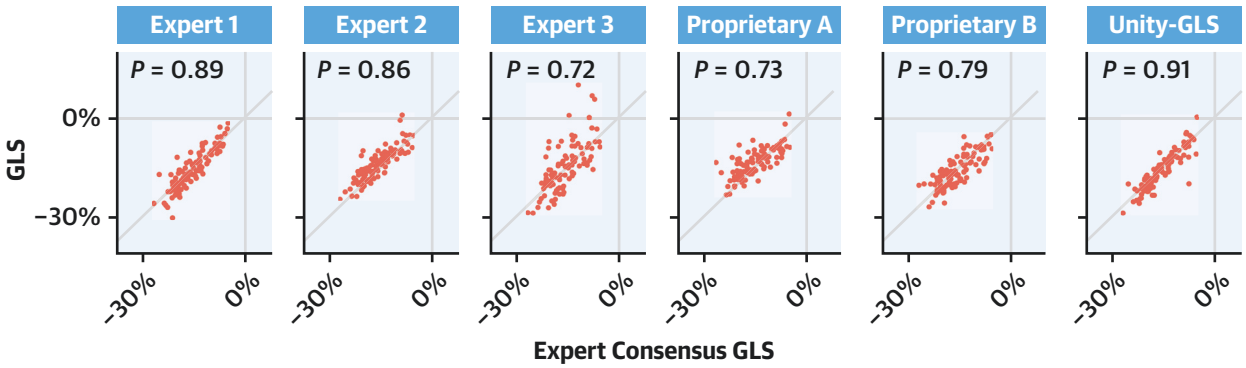
Each study contributes an advancement that can be built on by future workers. For example, Unity-GLS shares the property of Wei et al[16] of using multiple frames, although our use of the preceding and subsequent frames is simply to supply collateral

CENTRAL ILLUSTRATION **Multiexpert Consensus Defines the Reference Standard**



We use 11 expert opinions, in 2 ways:
Their **mean** (white) is the expert consensus reference,
Their individual **variation** defines how closely the AI (blue) must match the consensus, to be good enough.





Stowell CC, et al. JACC Cardiovasc Imaging. 2024;17(8):865–876.

Individual expert opinions, and individual machine measurements can be compared with the expert consensus as the reference standard. Individual experts E1 to E11 showed correlations with the expert consensus ranging from 0.67 to 0.89. proprietary softwares showed correlations of 0.73 and 0.79. The UK Unity open-source network showed correlation of 0.91. All $P < 0.001$. AI = artificial intelligence; GLS = global longitudinal strain.

temporal information, which enabled an improvement beyond our previous single-frame network experiments. A common drive for many studies has been starvation of expertly labelled data. Some groups have tackled this issue by generating synthetic images.[18-21] Our study and associated data should help to rectify this. It also provides apical 4-, 2-, and 3-chamber images, which are needed for GLS, and covers a range of vendors.

PROPRIETARY MEASUREMENTS OF GLS VIA SPECKLE TRACKING. Proprietary method B was introduced as a replacement for proprietary method A by the same vendor, presumably intended as an improvement. The correlation between the 2 proprietary measurements was 0.77. There was no evidence in our study of a substantial improvement in performance, either against the consensus of expert GLS measurements or against their visual assessments of longitudinal function. It is difficult for researchers to ascertain what changed between the 2 or how any aspect of their performance might be improved because the steps in their processes are confidential.

ESTABLISHING A GOLD STANDARD WHEN EXPERTS ARE NOT UNANIMOUS. When training a clinician to make measurements, it is recognized that there are

some measurements for which all experts agree closely: the trainee must answer similarly to be considered competent. In contrast, there are other measurements for which experts differ in their judgment and the trainee may, therefore, be permitted a wider range of permissible values.

We apply this principle in 3 ways in this study. First, the reference standard comes from not a single expert, but the consensus of several. This strategy improves the precision of the reference standard. Second, the spectrum of opinions of experts on each individual image effectively defines both the position and the size of the target the AI must hit. Third, cases with a wide spectrum of expert opinion are automatically identified as inherently difficult.

We have not attempted to use an alternative imaging modality method as the reference standard for several reasons. First, there would inevitably be differences in imaging plane between modalities resulting in variation that would seem to be an error, but was not. Second, we want the AI GLS to be explainable, that is, with the intermediate step of construction of endocardial curve from the echo image open to inspection so that the source of errors can be explored. Finally, we wanted the reference standard to mirror that of clinical trainees, in being the expert consensus interpretation of that image, and not a different image.

VALIDATION AGAINST THE EXPERT INTUITION OF LONGITUDINAL FUNCTION. Having a panel of experts also allowed us to establish whether the GLS measurements (human or machine) were indeed related to the intuitive concept of longitudinal function. We chose a ranking process that consisted of ranking each study multiple times against other studies. This process provides a relative assessment of LV function, which, with enough comparisons, can provide a high precision. This method is advantageous for several reasons. First, the granularity of ranking is greater than choosing a small number of categories with arbitrary distinctions (eg, mild vs moderate). Second, high granularity ratings do not require experts to assign values with unreasonable precision (eg, an ejection fraction of 34.5%). Third, because each assessment is done within a single operator, there is no need for an elaborate prior calibration process. Expert 1's opinion of study 1 is never directly compared against expert 2's opinion of study 2; only the relative rankings of 2 studies by the same expert are used.

STUDY LIMITATIONS. This study focused entirely on GLS and not on other indices of LV function. GLS was chosen because it could be applied across many forms of echocardiography, from departmental, portable, and even handheld devices.

This study did not test whether the GLS measurements made by the individual experts, their consensus, or the 3 machine methods predict clinical outcomes such as mortality. Previous studies have documented the prognostic power of LV GLS extensively.[22-24]

The number of expert observers is only a small proportion of the UK's clinical staff conducting echocardiograms. However, many are clinical leaders within their departments and within their research field generally, and are involved in guideline development processes. The size of this group is already much larger than that of any other multiexpert assessment of GLS in the published reports.

When comparing the performance of machine methods vs human experts, it should be remembered that the human expert opinions contributed to the expert consensus standard. Therefore, the individual expert errors are biased to be slightly smaller, and slightly biased toward the experts and against the machine algorithms. Despite this factor, the machine methods performed adequately.

Our method of calculating GLS works from the length of the complete contour, and not from regional tracking of speckles. This is because we want our method to be verifiable at each step. A consequence of our method is that it will not provide information on regional strain, only global. Furthermore, because neither the experts nor the AI were constrained to ensure that the diastolic length was longer than the systolic length, in a handful of cases where there was very poor LV function and poor image quality, some of the strain values were positive.

This study aimed to test performance against human experts in their optimized state, that is, with a consistent single frame. The neural network similarly focused on a single frame, although with surrounding frames for context. However, future improvements might include the evaluation of intermediate timesteps seeking a temporally consistent curve of strain against time.[14]

CONCLUSIONS

An AI algorithm can be trained by a nationwide panel of experts to make clinical measurements fully automatically to an acceptable standard, that is, performing better than individual experts. A fully open method, with reproducible training, and a public validation process with full disclosure of graphical

results, may be an advisable step to help clinicians decide for themselves whether machine performance in image interpretation is clinically adequate.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Rajani has received speaker fees from Siemens Healthcare and GE Medical; and has provided consultancy to Medtronic and Edwards Lifesciences. Dr Rana has provided consultancy to Philips and Occlutech. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Darrel P. Francis, National Heart and Lung Institute, Imperial College London, B Block, 2nd Floor, Hammersmith Hospital, London W12 0HS, United Kingdom. E-mail: d.francis@imperial.ac.uk.

PERSPECTIVES

COMPETENCY IN MEDICAL KNOWLEDGE: GLS is an important prognostic tool, reported to be more reproducible and prognostic than ejection fraction. Despite this, it is not in widespread clinical use. Greater transparency with open methods may increase trust and uptake among clinicians and echocardiographers.

TRANSLATIONAL OUTLOOK: We provide 2 open-source trained algorithms that have been developed by a collaborative group of clinicians. We invite international researchers to build on our work, harnessing the code and data, which we provide open-source and free for reuse.

REFERENCES

1. Zhu D, Ito S, Miranda WR, Nkomo VT, Pislaru SV, Villarraga HR, et al. Left ventricular global longitudinal strain is associated with long-term outcomes in moderate aortic stenosis. *Circ Cardiovasc Imaging*. 2020;13:e009958.

2. Oikonomou EK, Kokkinidis DG, Kampaktsis PN, Amir EA, Marwick TH, Gupta D, et al. Assessment of prognostic value of left ventricular global longitudinal strain for early prediction of chemotherapy-induced cardiotoxicity: a systematic review and meta-analysis. *JAMA Cardiol*. 2019;4:1007–1018.

3. Liou K, Negishi K, Ho S, Russell EA, Cranney G, Ooi SY. Detection of obstructive coronary artery disease using peak systolic global longitudinal strain derived by two-dimensional speckle-tracking: a systematic review and meta-analysis. *J Am Soc Echocardiogr*. 2016;29:724–735.e4.

4. Bertini M, Ng ACT, Antoni ML, Nucifora G, Ewe SH, Auger D, et al. Global longitudinal strain predicts long-term survival in patients with chronic ischemic cardiomyopathy. *Circ Cardiovasc Imaging*. 2012;5:383–391.

5. Amzulescu MS, De Craene M, Langet H, Pasquet A, Vancraeynest D, Pouleur AC, et al. Myocardial strain imaging: review of general principles, validation, and sources of discrepancies. *Eur Heart J Cardiovasc Imaging*. 2019;20:605–619.

6. Orde S, Huang SJ, McLean AS. Speckle tracking echocardiography in the critically ill: enticing research with minimal clinical practicality or the answer to non-invasive cardiac assessment? *Anaesth Intensive Care*. 2016;44:542–551.

7. Howard JP, Stowell CC, Cole GD, Ananthan K, Demetrescu CD, Pearce K, et al. Automated left ventricular dimension assessment using artificial intelligence developed and validated by a UK-wide collaborative. *Circ Cardiovasc Imaging*. 2021;14:e011951.

8. Unity Imaging Collaborative - Open-Access Datasets for AI in Cardiology. Accessed November 21, 2023. https://data.unityimaging.net/

9. Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. *PloS One*. 2015;10:e0121945.

10. Hotelling H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann Math Stat*. 1940;11:271–283.

11. Computing R. *Core Team: R Foundation for Statistical Analysis*. R: a language and environment for statistical computing; 2020. https://www.R-project.org/

12. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging*. 2020;13:2017–2035.

13. Støylen A, Mølmen HE, Dalen H. Left ventricular global strains by linear measurements in three dimensions: interrelations and relations to age, gender and body size in the HUNT Study. *Open Heart*. 2019;6:e001050.

14. Smistad E, Østvik A, Haugen BO, Lovstakken L. *2D left ventricle segmentation using deep learning*. Washington, DC, USA: Paper presented at: 2017 IEEE International Ultrasonics Symposium (IUS); September 6-9, 2017. https://doi.org/10.1109/ULTSYM.2017.8092812

15. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging*. 2019;38:2198–2210.

16. Wei H, Cao H, Cao Y, Zhou Y, Xue W, Ni D, et al. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. Paper presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference. Peru: Lima; October 4-8, 2020. https://doi.org/10.1007/978-3-030-59713-9_60

17. Painchaud N, Skandarani Y, Judge T, Bernard O, Lalande A, Jodoin PM. Cardiac segmentation with strong anatomical guarantees. *IEEE Trans Med Imaging*. 2020;39:3703–3713.

18. Duchateau N, Sermesant M, Delingette H, Ayache N. Model-based generation of large databases of cardiac images: synthesis of pathological cine mr sequences from real healthy cases. *IEEE Trans Med Imaging*. 2018;37:755–766.

19. Zhou Y, Giffard-Roisin S, De Craene M, Camarasu-Pop S, D'Hooge J, Alessandrini M, et al. A framework for the generation of realistic synthetic cardiac ultrasound and magnetic resonance imaging sequences from the same virtual patients. *IEEE Trans Med Imaging*. 2018;37:741–754.

20. Alessandrini M, Heyde B, Queiros S, Cygan S, Zontak M, Somphone O, et al. Detailed evaluation of five 3D speckle tracking algorithms using synthetic echocardiographic recordings. *IEEE Trans Med Imaging*. 2016;35:1915–1926.

21. D'hooge J, Barbosa D, Gao H, Claus P, Prater D, Hamilton J, et al. Two-dimensional speckle tracking echocardiography: standardization efforts based on synthetic ultrasound data. *Eur Heart J Cardiovasc Imaging*. 2016;17:693–701.

22. Biering-Sørensen T, Biering-Sørensen SR, Olsen FJ, Sengeløv M, Jørgensen PG, Mogelvang R, et al. Global Longitudinal Strain by Echocardiography predicts long-term risk of cardiovascular

morbidity and mortality in a low-risk general population: the Copenhagen City Heart Study. *Circ Cardiovasc Imaging*. 2017;10:e005521.

23. Magne J, Cosyns B, Popescu BA, Carstensen HG, Dahl J, Desai MY, et al. Distribution and prognostic significance of left ventricular global longitudinal strain in asymptomatic significant aortic stenosis: an individual participant data meta-analysis. *JACC Cardiovasc Imaging*. 2019;12:84–92.

24. Tower-Rader A, Mohananey D, To A, Lever HM, Popovic ZB, Desai MY. Prognostic value of global longitudinal strain in hypertrophic cardiomyopathy: a systematic review of existing literature. *JACC Cardiovasc Imaging*. 2019;12:1930–1942.

KEY WORDS artificial intelligence, global longitudinal strain, echocardiography

APPENDIX For an expanded Methods section as well as supplemental tables, figures, and references, please see the online version of this paper.