# Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test–retest performance

*João B Augusto, Rhodri H Davies, Anish N Bhuva, Kristopher D Knott, Andreas Seraphim, Mashael Alfarih, Clement Lau, Rebecca K Hughes, Luís R Lopes, Hunain Shiwani, Thomas A Treibel, Bernhard L Gerber, Christian Hamilton-Craig, Ntobeko A B Ntusi, Gianluca Pontone, Milind Y Desai, John P Greenwood, Peter P Swoboda, Gabriella Captur, João Cavalcante, Chiara Bucciarelli-Ducci, Steffen E Petersen, Erik Schelbert, Charlotte Manisty, James C Moon*

## Summary

**Background** Left ventricular maximum wall thickness (MWT) is central to diagnosis and risk stratification of hypertrophic cardiomyopathy, but human measurement is prone to variability. We developed an automated machine learning algorithm for MWT measurement and compared precision (reproducibility) with that of 11 international experts, using a dataset of patients with hypertrophic cardiomyopathy.

**Methods** 60 adult patients with hypertrophic cardiomyopathy, including those carrying hypertrophic cardiomyopathy gene mutations, were recruited at three institutes in the UK from August, 2018, to September, 2019: Barts Heart Centre, University College London Hospital (The Heart Hospital), and Leeds Teaching Hospitals NHS Trust. Participants had two cardiovascular magnetic resonance scans (test and retest) on the same day, ensuring no biological variability, using four cardiac MRI scanner models represented across two manufacturers and two field strengths. End-diastolic short-axis MWT was measured in test and retest by 11 international experts (from nine centres in six countries) and an automated machine learning method, which was trained to segment endocardial and epicardial contours on an independent, multicentre, multidisease dataset of 1923 patients. Machine learning MWT measurement was done with a method based on solving Laplace's equation. To assess test–retest reproducibility, we estimated the absolute test–retest MWT difference (precision), the coefficient of variation (CoV) for duplicate measurements, and the number of patients reclassified between test and retest according to different thresholds (MWT >15 mm and >30 mm). We calculated the sample size required to detect a prespecified MWT change between pairs of scans for machine learning and each expert.

**Findings** 1440 MWT measurements were analysed, corresponding to two scans from 60 participants by 12 observers (11 experts and machine learning). Experts differed in the MWT they measured, ranging from 14·9 mm (SD 4·2) to 19·0 mm (4·7; p<0·0001 for trend). Machine learning-measured mean MWT was 16·8 mm (4·1). Machine learning precision was superior, with a test–retest difference of 0·7 mm (0·6) compared with experts, who ranged from 1·1 mm (0·9) to 3·7 mm (2·0; p values for machine learning *vs* expert comparison ranging from <0·0001 to 0·0073) and a significantly lower CoV than for all experts (4·3% [95% CI 3·3–5·1] *vs* 5·7–12·1% across experts). On average, 38 (64%) patients were designated as having MWT greater than 15 mm by machine learning compared with 27 (45%) to 50 (83%) patients by experts; five (8%) patients were reclassified in test–retest by machine learning compared with four (7%) to 12 (20%) by experts. With a cutoff point of more than 30 mm for implantable cardioverter-defibrillator, three experts would have changed recommendations between tests a total of four times, but machine learning was consistent. Using machine learning, a clinical trial to detect a 2 mm MWT change would need 2·3 times (range 1·6–4·6) fewer patients.

**Interpretation** In this preliminary study, machine learning MWT measurement in hypertrophic cardiomyopathy is superior to human experts with potential implications for diagnosis, risk stratification, and clinical trials.

**Funding** European Regional Development Fund and Barts Charity.



## Introduction

Left ventricular maximum wall thickness (MWT) is a key imaging biomarker in hypertrophic cardiomyopathy, guiding diagnosis, risk stratification, and clinical management.[1–4] For diagnosis, hypertrophic cardiomyopathy is clinically defined by an MWT of at least 15 mm in one or more left ventricular myocardial segments in the absence of abnormal loading conditions,

*Lancet Digit Health* 2020

Published **Online** December 3, 2020 https://doi.org/10.1016/S2589-7500(20)30267-3

**Cardiac Imaging Department, Barts Heart Centre, St Bartholomew's Hospital, London, UK** (J B Augusto MD, R H Davies PhD, A N Bhuva PhD, K D Knott MBBS, A Seraphim MBBS, M Alfarih MSc, C Lau MBChB, R K Hughes MBBS, L R Lopes PhD, H Shiwani MBBS, T A Treibel PhD, S E Petersen PhD, C Manisty PhD, J C Moon MD); **Institute of Cardiovascular Science, University College London, London, UK** (J B Augusto, R H Davies, A N Bhuva, K D Knott, A Seraphim, M Alfarih, R K Hughes, L R Lopes, T A Treibel, G Captur PhD, C Manisty, J C Moon); **William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London, UK** (C Lau, S E Petersen); **Division of Cardiology, Department of Cardiovascular Diseases, Cliniques Universitaires St Luc UCL, Woluwe St Lambert, Belgium** (B L Gerber PhD); **Pôle de Recherche Cardiovasculaire, Institut de Recherche Expérimentale et Clinique, Université Catholique de Louvain, Brussels, Belgium** (B L Gerber); **The Prince Charles Hospital, Brisbane, QLD, Australia** (C Hamilton-Craig PhD); **Centre for Advanced Imaging, University of Queensland and Griffith University School of Medicine, QLD, Australia** (C Hamilton-Craig); **Division of Cardiology, Department of Medicine, University of**

Cape Town and Groote Schuur Hospital, Cape Town, South Africa (N A B Ntusi DPhil); Hatter Institute of Cardiovascular Research in Africa and Cape Universities Body Imaging Centre, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa (N A B Ntusi); Department of Cardiovascular Imaging, Centro Cardiologico Monzino IRCCS, Milan, Italy (G Pontone PhD); Heart and Vascular Institute Cleveland Clinic, Cleveland, OH, USA (M Y Desai MD); Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, and Leeds Teaching Hospitals NHS Trust, UK (J P Greenwood PhD, P P Swoboda PhD); Minneapolis Heart Institute, Department of Cardiology, Abbott Northwestern Hospital, Minneapolis, MN, USA (J Cavalcante MD); Valve Science Center, Minneapolis Heart Institute Foundation, Minneapolis, MN, USA (J Cavalcante); Bristol Heart Institute, Bristol National Institute of Health Research Biomedical Research Centre, University Hospitals Bristol NHS Trust and University of Bristol, Bristol, UK (C Bucciarelli-Ducci PhD); Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA (E Schelbert MD); Cardiovascular Magnetic Resonance Center, UPMC Heart and Vascular Institute, University of Pittsburgh Medical Center, Pittsburgh, PA, USA (E Schelbert)

Correspondence to:
Prof James Moon, St Bartholomew's Hospital, West Smithfield, London EC1A 7BE, UK
**j.moon@ucl.ac.uk**

## Research in context

### Evidence before this study

Both diagnosis and risk stratification of hypertrophic cardiomyopathy relies on accurate measurement of left ventricular maximum wall thickness (MWT), but its measurement by human experts has variation. Machine learning approaches with fully convolutional neural networks could improve precision in MWT measurement. We searched PubMed for studies published before Aug 7, 2020, focused on automated left ventricular wall thickness measurement, using the terms ("left ventricle" OR "LV") AND ("maximum wall thickness" OR "MWT" OR "wall thickness") AND ("automated" OR "automatic"), without language restrictions. This process returned 53 publications. The articles covered a range of modalities (echocardiogram, cardiac MRI, cardiac CT, and PET) and tools (manual, semi-automated or fully automated, two dimensional or three dimensional imaging, different sequences and parameters). In most cases, the aim was to compare left ventricular segmentation tools between different modalities with a particular focus on left ventricular volumes, mass, or measures of myocardial deformation. Some tools required manual endocardial and epicardial left ventricular segmentation before automatic MWT measurement. To the best of our knowledge, the precision of deep learning MWT measurement when compared with manual segmentation by human experts is still unknown.

### Added value of this study

MWT measurement lacks standardisation. Here, we propose a machine learning solution with superior precision to human experts. Our automated machine learning method, trained on 1923 separate multicentre, multidisease cases, segments the endocardial and epicardial left ventricular borders and then measures left ventricular MWT, substantially surpassing the MWT measurement performance of 11 tested human experts, gauged by precision. Our machine learning tool was more consistent than human experts in assigning diagnoses of hypertrophic cardiomyopathy (MWT >15 mm) and recommendations for an implantable cardioverter-defibrillator (ICD) and returned smaller changes in sudden cardiac death risk score. The precision shown by our machine learning tool is the first step in improving MWT measurement.

### Implications of all the available evidence

Machine learning holds the promise of transforming health care. Widespread adoption of such a tool to measure MWT could affect clinical decision making in hypertrophic cardiomyopathy by improving precision in diagnosis and risk stratification for ICD, while also reducing sample sizes needed for clinical trials that use MWT as an outcome measure. Our tool could also easily be applied retrospectively (and prospectively) to pivotal clinical trials in cardiology and potentially change research outcomes. A tool that is more precise and has the same accuracy as human experts could be the next gold standard.

with a lower threshold in familial disease.[3] MWT is also used in decision making for primary prevention of sudden cardiac death with an implantable cardioverter-defibrillator (ICD).[1,3] An example of the latter is the hypertrophic cardiomyopathy sudden cardiac death risk score (HCM Risk-SCD) recommended by the European Society of Cardiology, which uses MWT as a continuous variable to stratify risk in these patients, while the American Heart Association suggests an MWT cutoff of more than 30 mm for ICD.[1,3,5]

Measurement error therefore can lead to under-diagnosis and over-diagnosis, as well as inappropriate or ineffective therapies, yet there is no standardised protocol for measuring MWT. Sources of MWT measurement error include the complex myocardial shape (trabeculation, non-parallel three-dimensional edges), modality-specific variation (spatiotemporal resolution, piloting, artifacts) and intra-observer and interobserver variation.[6] Cardiovascular magnetic resonance (CMR) can offer better spatial resolution and blood–myocardium interface definition,[7–10] while automated machine learning approaches could minimise human variation.[9,10] In many domains, machine learning approaches with deep fully convolutional neural networks (CNNs) have achieved human performance,[11,12] including in cardiac imaging.[10] The definition of a gold standard typically relies on comparisons across modalities (with systematic bias) and measurements by experts (with intra-observer and interobserver variability). Ideally, these comparisons would be made against a ground truth (for accuracy), but for MWT, a ground truth is not easily attainable. One solution to objectively assess performance in MWT measurement is to evaluate repeatability (ie, precision) on test–retest data—ie, multiple scans taken from the same individual.[10]

We applied a machine learning algorithm previously trained on multicentre, multidisease cases to deliver automated left ventricular contours and MWT on a separate, multicentre test–retest dataset of hypertrophic cardiomyopathy CMR scans. We aimed to investigate whether MWT by machine learning was more precise than a multicentre panel of 11 human experts (nine centres in six countries, across four continents) by evaluating test–retest MWT measurements, and to explore the implications of machine learning MWT measurements on hypertrophic cardiomyopathy diagnosis and risk stratification.

## Methods

### Study population

For the hypertrophic cardiomyopathy test–retest dataset, patients with hypertrophic cardiomyopathy (including

seven genotype-positive, phenotype-negative patients with left ventricular MWT less than 15 mm, as clinically defined before enrolment) who underwent CMR for clinical reasons were opportunistically recruited at three institutions in the UK from August, 2018, to September, 2019: Barts Heart Centre, University College London Hospital (The Heart Hospital), and Leeds Teaching Hospitals NHS Trust. The seven phenotype-negative patients were included so as to assess diagnosis reclassification according to the 15 mm cutoff. Four scanner models (Siemens Aera, Philips Achieva, Siemens Avanto, Siemens Prisma) are thus represented across two MRI scanner manufacturers (Siemens Healthineers [Erlangen, Germany], Philips Healthcare [Amsterdam, Netherlands]) and two field strengths (1·5 and 3 Tesla). Exclusion criteria were patients younger than 18 years, contraindications to CMR, cardiac implantable electronic devices, clinically significant arrhythmia (eg, atrial fibrillation, frequent ectopy) or inability to hold breath, and pregnancy. Ethical approval was obtained in each centre (London–Surrey Research Ethics Committee, reference 18/LO/0188) and the study conformed to the principles of the Helsinki Declaration. Written informed consent was obtained from all participants.

### CMR scan protocol

All patients underwent CMR scans twice in the same day (scans A and B). After the first scan, patients were brought out of the bore, repositioned on the table, and isocentre positioning was repeated before the second scan. Each scan used a similar protocol (without changes in imaging parameters between tests) that consisted of balanced steady-state free precession (bSSFP) cine imaging in four-chamber, two-chamber, and three-chamber views and two-dimensional (2D) left ventricular short-axis (SAX) cine stack, ensuring coverage of the left ventricular base and apex, as per international recommendations.[8] Cine imaging was acquired in both scans before any gadolinium-based contrast was given. CMR parameters are detailed in the appendix (p 9). Image quality was assessed using previously published criteria.[13] All cines used in this study had an overall score no greater than 3 (on a scale from 0 to 21, with lower scores indicating higher quality images) as judged by an investigator (JBA), as detailed in the appendix (p 8). Of note, all scans scored 0 for the left ventricular coverage criteria, indicating full left ventricular coverage from base to apex.[13]

### Machine learning algorithm

We previously developed an automated 2D deep fully CNN[14] with U-net architecture[15] that was trained to segment the left ventricular SAX endocardial and epicardial contours in end-diastole and end-systole from an input of CMR bSSFP cine images.[10] In brief, the CNN was trained on an independent dataset of CMR scans from 1923 patients, with each scan comprising three sets of standard cine images: two-chamber and four-chamber views and a stack of 2D SAX slices (mean 12 slices), each with 25 or 30 frames. These images were manually annotated by an expert (JCM). Manual segmentation in the SAX stack consisted of left ventricular endocardial and epicardial contours in end-diastole and left ventricular endocardial contours in end-systole. Papillary muscles and trabeculations were considered part of the left ventricular blood pool. The training dataset comprised healthy volunteers and individuals with balanced pathologies (diseases with dilatation and hypertrophy, the latter including hypertrophic cardiomyopathy; Fabry disease; hypertension; amyloidosis; and aortic stenosis) who were scanned across 13 centres in three countries, using a variety of scanners (ten scanner models, 1·5 and 3 Tesla field strengths, and three MRI manufacturers: Siemens Healthineers, Philips Healthcare, GE Healthcare [Chicago, IL, USA]). We previously validated our machine learning algorithm on the VOLUMES Resource dataset, a precision open-source dataset for left ventricular volumes and mass.[10] Further technical details of the machine learning training method are available in the appendix (pp 2–6).

For the **VOLUMES Resource dataset** see https://thevolumesresource.com/

### MWT measurement

MWT measurement by machine learning proceeded in two steps: drawing of left ventricular endocardial and epicardial contours and wall thickness measurement. Endocardial left ventricular contours were automatically drawn on each SAX cine image using the machine learning algorithm (appendix pp 2–6). Each 2D SAX image was segmented separately and the contours for each phase interpolated across all slices to calculate the volume for that phase. End-diastole was chosen as the phase with the largest blood pool volume and both endocardial and epicardial contours were drawn in this phase. Further details of this step can be found in the appendix (pp 4–6).

Measurement of the distance between epicardial and endocardial contours is ambiguous and commonly used methods have limitations (appendix p 5). Desirable characteristics of an effective approach include a unique and invertible solution (ie, homeomorphic mapping), such that each point on the endocardial contour maps to a single corresponding epicardial point and vice versa; being consistent and robust, such that small changes in contour shape do not lead to large changes in measurement; and being computationally efficient. There is no accepted method to measure MWT. A previous study suggested the use of Laplace's equation to measure cerebral cortical thickness in neuroimaging.[16] We used a similar method, solving Laplace's equation to define a dense correspondence between myocardial borders and the maximum distance recorded (figure 1; appendix pp 4–5).

See **Online** for appendix

The largest wall thickness value across all slices for each scan was recorded as the MWT. Image files were reviewed for quality control but there was no active human correction of any of the machine learning MWT

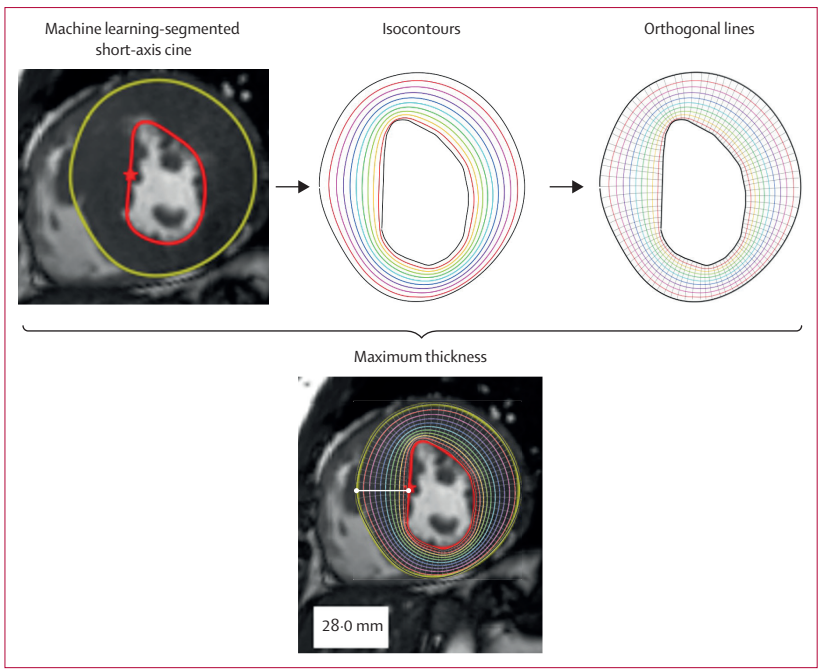


***Figure 1:*** **A solution of Laplace's equation for left ventricular wall thickness measurement**
First, automated machine learning end-diastolic contours are obtained and used to specify boundary conditions. Laplace's equation is solved for the whole field and shown here using isocontours (different colours). The gradient field can be followed from endocardium (red) to epicardium (yellow), generating a set of non-intersecting streamlines that are orthogonal to the isocontours. The wall thickness corresponds to the Euclidian distance between the start and end points of the streamline (resulting in no overlap between wall thickness calliper positions). Here, the maximum thickness measures 28·0 mm.

measurements—these would either be accepted or rejected. If there was a Turing test failure in any image (biologically implausible contouring) as deemed by JBA (investigator with Level 3 CMR certification), that image was excluded. All other images were included for final analysis.

11 CMR and cardiomyopathy experts (BLG, CB-D, CH-C, CM, ES, GP, JC, JCM, MYD, NABN, and SEP), who were accredited as Level 3 by the Society for Cardiovascular Magnetic Resonance or European Association of Cardiovascular Imaging or who had at least 10 years of CMR experience, measured MWT in the hypertrophic cardiomyopathy test–retest dataset. Experts were aware that the sample was comprised of clinical and subclinical hypertrophic cardiomyopathies, but were blinded to any other clinical characteristics. Each scan was assigned a randomly generated identification code so that scans were analysed in a random order for each batch (one batch for scan A and one for scan B), and all observers were blinded to test and retest status. Each scan included three long-axis views (four chamber, two chamber, and three chamber) and a SAX stack cine. Similar instructions for MWT measurement were given to all experts. Experts were instructed to measure the MWT using digital callipers on the SAX stack only (although they could use long-axis views to plan this measurement), in end-diastole, as routinely done in clinical practice and in accordance with international recommendations.[3] All experts used cvi42 software (version 5.9.x) for the purpose of this analysis (Circle Cardiovascular Imaging, Calgary, Canada). Once the first batch of scans (ie, the experts' test batch) was analysed, experts deleted those scans and only then would the second batch of scans (ie, the experts' retest batch) be given to them. There was a minimum of 24 h between the analysis of each batch. The calliper locations of all measurements were exported to visualise the source of measurement variation.

### Statistical analysis

Discrete variables are presented as absolute frequencies with percentages and continuous variables as mean (SD) or mean (95% CI). Interobserver agreement for MWT for all observers was assessed using the intraclass correlation coefficient (ICC; absolute agreement between single measures) and Lin's concordance correlation coefficient (CCC; agreement between two measures) for pairwise comparisons.[17] CCC was categorised as almost perfect (>0·99), substantial (0·95 to 0·99), moderate (0·90 to <0·95), or poor (<0·90). To assess test–retest reproducibility, we estimated the absolute MWT difference (precision) and the coefficient of variation (CoV) for duplicate measurements[18] using the root mean square method.[19,20] We also estimated the Bland-Altman bias and limits of agreement; these limits of agreement are a measure of precision, calculated as (mean difference between test and retest) $\pm 1{\cdot}96 \times$ SD, where a smaller range between these two limits indicates better precision. Paired Student's *t* test was used to compare absolute MWT test–retest differences and mean test–retest MWTs between machine learning and each expert. We compared test–retest precision between different observers using a linear mixed-effects regression model, with observers (including machine learning), test–retest category, and the interaction term of observers and test–retest category included as fixed effects and patients as a random effect (appendix p 7). MWT correlations using Pearson's correlation (*r*) and linear regressions with coefficients of determination ($R^2$) and 95% CIs were done between the first and second scans for each observer.

The sample size required (number of patients) to detect a prespecified MWT change between pairs of scans was calculated ($\beta$=0·90, $\alpha$=0·05). As a measure of reliability, the minimum detectable change—ie, the smallest detectable difference that is not due to random variation—was estimated as $1{\cdot}96 \times \sqrt{2} \times \text{SD} \times \sqrt{(1-\text{ICC})}$.[21]

The proportion of patients reclassified between test and retest according to the 15 mm cutoff (for diagnosis) and 30 mm cutoff (for prognosis) are presented. We also assessed how much changes in MWT measurement between test and retest would affect the risk score generated by the HCM Risk-SCD calculator.[1] The MWT weighting in the HCM Risk-SCD follows a quadratic function[1] (appendix p 7), with results

expressed as a percentage risk of sudden cardiac death at 5 years. We noted the highest errors in MWT score weighting between test and retest. Two-sided p<0·05 was considered significant. All analyses were done in R (RStudio version 1.1.423). Full details are included in the appendix (pp 6–7).

### Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the paper for publication. JBA, RHD, and JCM had direct access and verified all the data in the study, and had final responsibility for the decision to submit for publication.

## Results

Scans were obtained from 60 patients with hypertrophic cardiomyopathy, with a mean age of 55·9 years (SD 13·4), 42 (70%) of whom were male (appendix p 10). A total of 1439 SAX images were included. Two scans per patient provided a dataset of 1440 MWT measurements across the 12 observers (11 experts and machine learning) for subsequent analysis. Both test and retest scans were available for analysis in all patients. Human quality control of machine learning wall thickness measurements revealed eight (0·6%) biologically implausible callipers out of the 1439 SAX images analysed (corresponding to eight images in five unique patients), all of which were excluded from our analysis, but none of the measurements represented an MWT, so this human quality control did not affect any results.

Across the patients analysed, mean MWT (between test and retest) for human experts ranged from 14·9 mm (SD 4·2) to 19·0 mm (4·7; p<0·0001 for trend), a difference of 4·1 mm. The machine learning-measured mean MWT was 16·8 mm (4·1), falling within the range measured by experts, with five experts measuring significantly thicker and four significantly thinner; two experts were statistically indistinguishable (table). The highest MWT difference noted between experts in a single case was 11·4 mm (MWT of 25·3 mm by expert 8 *vs* 36·7 mm by expert 9).

There was only moderate agreement among all experts for mean MWT test–retest (ICC 0·82, 95% CI 0·69–0·90). Pairwise agreement between experts was poor in two thirds of cases (CCC <0·90 in 37 [67%] of 55 pairs). A similarly poor pairwise agreement rate was also found between seven (64%) experts and machine learning.

When assessing test–retest reproducibility, experts had significant differences in precision (absolute test–retest MWT difference), ranging from 1·1 mm (SD 0·9) to 3·7 mm (2·0), with a mean difference across all experts of 1·5 mm (0·6; table). Test–retest difference among experts increased with higher mean MWT values ($r$=0·55; p<0·0001), but this trend was not seen with machine learning ($r$=0·04; p=0·783; appendix p 14).

Machine learning performance surpassed humans on all measures of test–retest error: its precision was significantly better than that of all experts (table; p values for machine learning *vs* expert comparison ranging from <0·0001 to 0·0073); the Bland-Altman limits of agreement were narrower for machine learning, with an interval half that of the human experts' mean (3·7 mm *vs* 7·7 mm; table; figure 2; appendix pp 15–17); the CoV for duplicate measurements was significantly lower in machine learning than for all experts (table); the coefficient of determination for the linear regression model of test MWT predicting retest MWT was higher in machine learning ($R^2$=0·96) than for all experts (ranging from 0·58 to 0·93; appendix pp 15–17); and machine learning did not show a significant contribution to MWT test–retest difference in a linear mixed-effects model (appendix p 11).

Using the 15 mm cutoff for hypertrophic cardiomyopathy diagnosis, five (8%) patients assessed by machine learning would have had a different diagnosis between test and retest scans, whereas human experts would have reclassified between four (7%) and 12 (20%) patients (mean eight patients [SD 3]), with only one expert reclassifying fewer patients than machine learning (appendix p 12). Among all scans (test and retest), the percentage of patients with hypertrophic cardiomyopathy diagnosis by machine learning was 64% on average (39 [65%] patients by test scan and 38 [63%] patients by retest scan), whereas among all experts, the range of hypertrophic cardiomyopathy diagnoses varied between 27 (45%) patients and 50 (83%) patients, which represents a 38% absolute difference (23 of 60 patients). For an individual patient, the minimal detectable change using machine learning was 0·4 mm, but varied between 1·0 mm and 3·8 mm with human expert analysis (appendix p 13).

| | Mean (SD) MWT, mm | Mean (SD) absolute MWT difference, mm | Bland-Altman bias, mm (limits of agreement) | Coefficient of variation (95% CI) | p value* |
|---|---|---|---|---|---|
| Machine learning | 16·8 (4·1) | 0·7 (0·6) | –0·1 (–2·0 to 1·7) | 4·3% (3·3 to 5·1) | .. |
| Expert 1 | 16·4 (1·9) | 3·7 (2·0)† | 0·4 (–5·0 to 5·9) | 12·1% (8·2 to 15·0) | <0·0001 |
| Expert 2 | 15·7 (3·9)† | 2·0 (1·7)† | –1·1 (–5·9 to 3·7) | 11·9% (8·8 to 14·4) | <0·0001 |
| Expert 3 | 19·0 (4·7)† | 1·6 (1·5)† | –0·2 (–4·5 to 4·1) | 7·9% (6·1 to 9·4) | <0·0001 |
| Expert 4 | 17·4 (3·9)‡ | 1·7 (1·4)† | 0·3 (–3·9 to 4·6) | 8·7% (7·2 to 10·0) | <0·0001 |
| Expert 5 | 14·9 (4·2)† | 1·9 (1·3)† | 1·1 (–2·9 to 5·1) | 11·2% (9·4 to 12·8) | <0·0001 |
| Expert 6 | 15·8 (4·2)† | 1·4 (1·2)† | –0·1 (–3·8 to 3·6) | 8·3% (6·5 to 9·8) | <0·0001 |
| Expert 7 | 18·9 (4·9)† | 1·4 (1·2)† | –0·1 (–3·7 to 3·4) | 6·7% (5·3 to 7·9) | 0·0006 |
| Expert 8 | 15·8 (4·1)† | 1·2 (1·2)‡ | –0·2 (–3·6 to 3·2) | 8·3% (5·6 to 10·3) | <0·0001 |
| Expert 9 | 18·6 (4·7)† | 1·3 (1·2)‡ | –0·8 (–4·0 to 2·4) | 6·9% (5·1 to 8·3) | 0·0003 |
| Expert 10 | 16·7 (3·9) | 1·1 (0·9)‡ | 0·0 (–2·8 to 2·8) | 5·7% (4·6 to 6·5) | 0·034 |
| Expert 11 | 19·0 (4·7)† | 1·3 (1·1)† | –1·0 (–3·7 to 1·7) | 5·9% (4·5 to 7·1) | 0·013 |
| Expert mean | 17·1 (4·1)‡ | 1·5 (0·6)† | –0·2 (–4·0 to 3·7) | .. | .. |

Mean MWT between test–retest is across all study participants. MWT=maximum wall thickness. *Comparing the coefficient of variance with that of machine learning. †p<0·001 *vs* machine learning using paired Student's *t* test. ‡p<0·05 *vs* machine learning using paired Student's *t* test.

***Table:*** **Test–retest reproducibility of MWT, by observer, and comparison with machine learning**

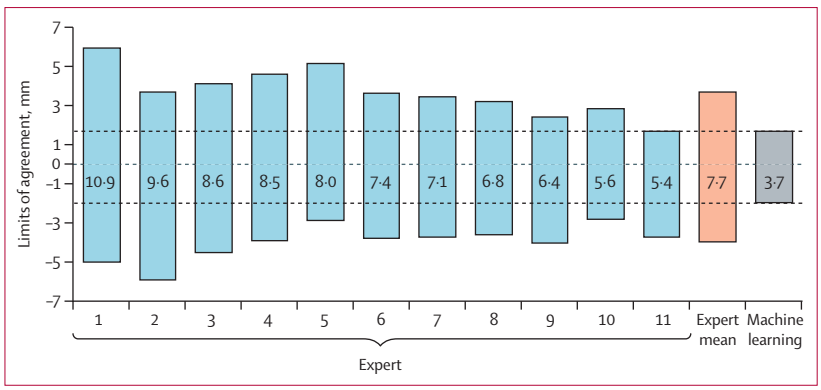


***Figure 2*: Bland-Altman limits of agreement intervals, by observer**
Limits of agreement are shown for each expert, the average of all 11 experts, and for machine learning. Bars represent the difference between the upper and lower limits (dashed lines represent the limits for machine learning) and are centred on the Bland-Altman bias (ie, mean difference between test and retest).

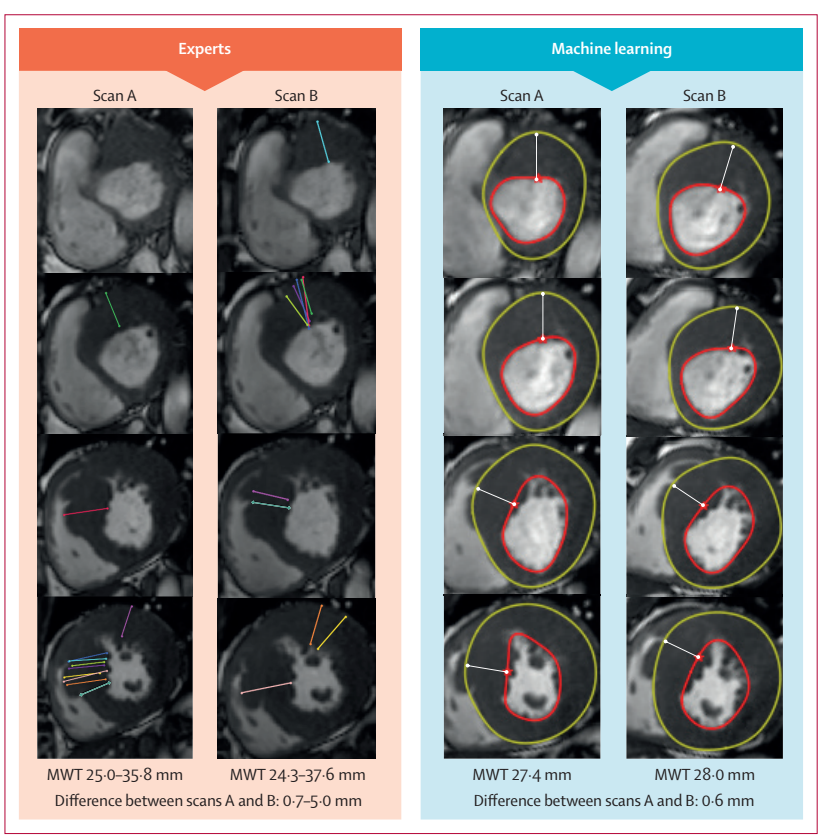


***Figure 3*: Example of extreme septal hypertrophy**
Machine learning picked phase 30 (out of 30 phases). Nine experts picked phase 1 and two experts picked phase 30 but, for illustration purposes, phase 1 is presented for experts. MWT=maximum wall thickness.

The maximum recorded difference in the HCM Risk-SCD due to MWT imprecision varied between 0·26 and 0·59 percentage points among experts, while the highest risk score error seen with machine learning was 0·19 percentage points, making machine learning up to 1·4–3·1 times more consistent (appendix p 19). If using the American Heart Association-recommended cutoff for ICD of more than 30 mm, two experts would have reclassified one patient, and one expert would have reclassified two patients between test and retest scans, whereas machine learning would have reclassified none (appendix p 12). Using this cutoff point, five of 11 human experts would have referred at least one patient for ICD but machine learning would have referred none (appendix p 12).

In our sample size assessments, we found that machine learning analysis of a clinical study would require considerably smaller sample sizes than human expert analysis for a prespecified MWT change and, as this prespecified change decreased, the difference between machine learning and the experts increased. Sample sizes per observer for different MWT changes are detailed in the appendix (p 13). For instance, sample size would be on average 2·3 times (range 1·6–4·6) smaller using machine learning to detect a 2 mm change in MWT (appendix p 13).

We present an example case of extreme left ventricular hypertrophy and discordance between experts due to exuberant trabeculation (figure 3). Within the same scan, experts picked different segments and locations in the myocardium and even different slices to measure MWT, leading to considerable disagreement in MWT measurement (10·8 mm difference in scan A and 13·3 mm difference in scan B). Between tests, some experts changed their measurement to a completely different location or slice, leading to remarkable test–retest differences (varying between 0·7 mm and 5·0 mm). Different degrees of trabeculae inclusion can be appreciated. Importantly, four experts would have recommended ICD for this patient in either test based on a 30 mm cutoff point, whereas seven experts and machine learning would not. One expert would have changed ICD recommendation between tests. Machine learning showed a smaller test–retest error than for all experts.

Frequency of error in MWT measurement cannot be formally determined among humans, as there is no definition of correct or incorrect measurements. We thus present a classification of the human variations in MWT measurement that can lead to discordance between experts, using examples from our study (figure 4). For machine learning, however, two main types of error can be identified: Turing test failures—ie, contours that are biologically implausible and can be easily identified by even non-experts (who have at least some familiarity with the modality)—and mis-segmentation (figure 4). Turing test failures were present in eight images, which were all excluded from the analysis. Mis-segmentation is similar to human variations and so its occurrences were still included in the precision analysis; one investigator (JBA) determined the presence of mis-segmentation in five (4%) of 120 scans (selected for image quality assessment) assessed by the machine learning algorithm, but identification of mis-segmentation can be subjective and thus we included these five scans in our analysis of machine learning precision.

## Discussion

Both diagnosis and risk stratification of hypertrophic cardiomyopathy relies on accurate MWT measurement by humans. However, this can be challenging. We present an automated machine learning method of measuring MWT in hypertrophic cardiomyopathy and have shown its superior precision against an international group of experts. Widespread adoption of such a tool could affect clinical decision making in hypertrophic cardiomyopathy by improving precision in diagnosis and risk stratification for ICD, while also reducing sample sizes needed for clinical trials that use MWT as an outcome measure. Further studies are needed to assess implementation in clinical practice.

MWT measurement by human experts is prone to under-diagnosis or over-diagnosis of hypertrophic cardiomyopathy. Our machine learning algorithm was able to diagnose hypertrophic cardiomyopathy on 64% of patients based on a 15 mm cutoff, right in the middle of the wide range of 45–83% diagnoses reported among experts. Importantly, up to one in five patients would have a different diagnosis between tests when assessed by experts (*vs* one in 12 using machine learning), highlighting important inconsistencies in routine analysis, with serious implications for clinical management and burden to the health-care system. Errors in sudden cardiac death risk stratification, on the other hand, could lead to inappropriate or ineffective therapies. There is substantial disagreement among experts, with nearly half of our experts recommending ICDs (using the 30 mm cutoff), three of whom changed their ICD recommendations between tests. By contrast, machine learning would maintain recommendation between tests and be more precise for sudden cardiac death risk stratification using the HCM Risk-SCD. Sample sizes to detect prespecified MWT interval changes in clinical trials would also decrease with machine learning. Furthermore, machine learning could easily be applied (either retrospectively or prospectively) to pivotal clinical trials in cardiology and potentially affect research results. However, although precision was thoroughly assessed in this study, the diagnostic and prognostic implications of this tool in clinical practice should be validated.

Machine learning and artificial intelligence are already being used in health care and are beginning to show their promised value. However, one of the main challenges is wide adoption in clinical practice. Our aim is to translate the algorithm into a tool that is routinely used in clinical care. We are currently investigating ways of achieving this, including in-line implementation directly on the magnetic resonance scanner and a cloud-based platform.

Humans instinctively follow a step-by-step method to measure MWT, but each step is prone to errors and variation. Accordingly, humans have to make multiple choices and identifications: the end-diastolic phase, the SAX slice with MWT (usually a quick visual scan is performed, but not all slices are measured), the thickest myocardial segment within a slice (again, not all segments are measured), two points in each myocardial border (which might result in different calliper angles between humans), and the calliper length itself (affected by problems in blood–myocardium interface definition, even in cases without overt left or right ventricle trabeculations; appendix p 18). Throughout these steps, there might also be clinical bias due to knowledge of the clinical status of the patient.

Interestingly, experts were inconsistent with each other, drawing MWT in different locations and even changing MWT locations in test–retest, highlighting how difficult it is to precisely measure MWT. It should be noted, however, that these human variations cannot be truly quantified as there is no formal definition of a correct MWT measurement. In our study, experts were allowed to screen all slices and phases for MWT, but very rarely was a calliper drawn in all slices; this could be a source of imprecision in humans, but also a reflection of

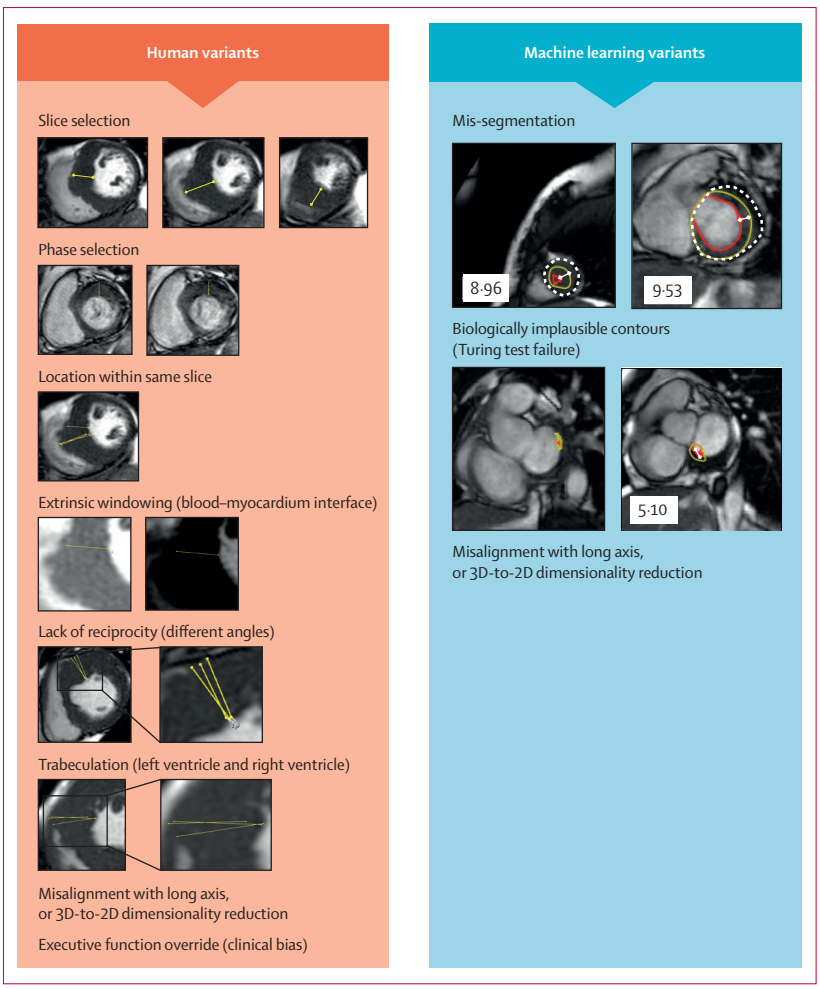


***Figure 4:*** **Variants seen in left ventricular wall thickness measurements**
2D=two dimensional. 3D=three dimensional.

what happens in the real world. The machine learning algorithm has a clear advantage over humans by scanning all slices and segments for the MWT measurement. What is more interesting, however, is the superior consistency (ie, precision) of the machine learning algorithm.

Several studies compare and validate segmentation tools against a single expert (or a small group). Here, we used a panel of 11 experts from four continents, to account for potential regional differences, and they showed a wide discrepancy in precision. It should be noted that our machine learning tool was overall superior not just to the average expert observer, but to all experts here studied. Furthermore, our tool was trained and validated in multiple scanners from multiple centres, making it more representative of the real world.

Other semi-automated tools can measure wall thickness, but these still rely on human corrections that are prone to variability. Machine learning is automated and more consistent than humans. First, the machine learning algorithm used is deterministic and thus intra-observer variability is absent (if one gives the algorithm the same image twice, it will give an identical solution). It should be said, however, that in this study the input was varied between test and retest, so the machine learning algorithm and humans were on equal ground. Second, it seems that the algorithm does not measure too much or too little as the average machine learning MWT lies at the midpoint of the experts' average range. Importantly, there is no established rule as to how clinicians should measure wall thickness. Here, we present a way of standardising wall thickness measurement using Laplace's equation.

Finally, in the absence of a ground truth, human errors cannot be formally defined and thus error rate cannot be determined. Nevertheless, we presented a classification of the variations that can be found in MWT measurement among humans. For machine learning, the algorithm produced a very low rate (0·6%) of biologically implausible contours.

We propose that the left ventricle should first be segmented by the machine learning tool in all cases. The machine learning output for each patient consists of an image file with the end-diastolic MWT in each slice. Humans (even those with minimal experience in cardiac imaging) can easily check if each of these calliper positions are biologically plausible and if they are (which we found to be true for 99% of the images corresponding to 96% of patients), then left ventricular MWT is noted. We suggest that if any of the segmentations are not plausible (Turing test failure), manual measurement should be performed instead. In this study, we opted to include the MWT measurements corresponding to the eight scans that had biologically implausible callipers, as the region of maximal thickness was clearly not affected, but we acknowledge that this could introduce variation in clinical practice. Notwithstanding, the expected precision yield in clinical practice from having to manually measure only 4% of cases would still be considerably high. It should be noted that the supposed mis-segmentations seen with machine learning can also be seen in humans and, as with human variations, it is difficult to ascertain exactly what segmentations are correct or incorrect. As such, these images were not excluded in the precision analysis.

We found scan–rescan variability to be lower for machine learning than for our 11 experts. Differences in test–retest would be zero by machine learning if the input given was the same (given the deterministic nature of machine learning), but this was not the case in this study because the images in our test and retest sets were fundamentally different due to variation in slice prescription by the radiographer between tests (see appendix p 4 for examples). We kept biological differences to a minimum (eg, haemodynamics) by doing test and retest scans sequentially on the same day.

Our study has several limitations. The machine learning tool is not fully automated, as humans still retain a quality control role. We expect implementation in clinical practice to be interactive: machine learning outputs image files with the left ventricular segmentation, which humans check for any Turing test failures. It might be possible to build an end-to-end neural network that can estimate MWT directly, without the need for segmentation, but we have chosen the approach described here since it is fully explainable, allowing clinicians to directly visualise how the thickness measurement was made, engendering trust and avoiding a black-box approach that clinicians, patients, and the public dislike when it concerns critical health issues. Only bSSFP cine images were used to train the CNN and some additional training images would be needed if we wished to use other cine sequences (eg, real time, gradient echo). Additionally, machine learning performance has not been tested in scans with poor image quality (eg, patients with inability to hold their breath, or patients with atrial fibrillation or implantable devices).

In conclusion, we have shown how an automated machine learning tool for MWT measurement in hypertrophic cardiomyopathy is feasible and more precise than an international group of experts. Precision is the first step in machine learning MWT measurement. The next step is to validate this tool for outcome prediction on a separate, ideally more heterogeneous, population—representing different centres (and thus multiple scanners and imaging parameters) and left ventricular hypertrophy patterns. This will provide effectiveness, safety and acceptability. A tool that is more precise, as shown here, and is at least non-inferior to human experts at predicting outcomes should, in principle, be the new gold standard. Widespread adoption of such a tool could help to improve hypertrophic cardiomyopathy diagnosis and sudden cardiac death risk stratification and substantially decrease sample sizes needed for clinical trials.

### Contributors

This study was conceived by JBA, RHD, LRL, CM, and JCM, with input from all other authors. JBA, ANB, KDK, AS, MA, CL, RKH, JPG, and PPS were involved in data collection. JBA, RHD, and JCM had direct access to and verified all the data in the study. JBA, RHD, ANB, HS, BLG, CH-C, NABN, GP, MYD, JC, CB-D, SEP, ES, CM, and JCM did the analyses with critical input from all other authors. JBA, RHD, LRL, TAT, BLG, CH-C, NABN, GP, MYD, JPG, PPS, GC, JC, CB-D, SEP, ES, CM, and JCM were involved in the writing of the manuscript. All authors were involved in data interpretation and critical revision of the manuscript. All authors approved the final version of the manuscript.

### Declaration of interests

SEP reports personal fees from Circle Cardiovascular Imaging, outside of the submitted work. JC reports research support from Siemens Healthineers and Circle Cardiovascular Imaging, outside of the submitted work. GP declares receiving honorarium as a speaker and research grants from GE Healthcare, Bracco, and Heartflow, outside of the submitted work. MYD is the principal investigator of the VALOR-HCM trial (NCT04349072), sponsored by Myokardia, outside of the submitted work. CB-D is the part-time Chief Executive Officer of the Society for Cardiovascular Magnetic Resonance. All other authors declare no competing interests.

### Data sharing

De-identified cardiac MRI scans and raw results included in this Article can be made available to researchers from accredited research institutions. Access to data is possible from the date of publication and will be limited to investigators who provide a methodologically sound proposal and can conduct analyses that achieve the aims of the proposal. Access to data requires a Material Transfer Agreement, which is examined and approved by the University College London Material Transfer Agreement team. The machine learning algorithm used in this manuscript is not widely available yet, but the authors are investigating ways of disseminating it. The authors, however, agree to apply the machine learning algorithm to data provided by other academic researchers on their behalf for research purposes only, following completion of a Material Transfer Agreement. Proposals and requests for data access should be directed to the corresponding author via email.

### Acknowledgments

RHD was funded through the CAP-AI programme by a grant from the European Regional Development Fund and Barts Charity. AS declares a doctoral research fellowship from the British Heart Foundation (FS/18/83/34025). LRL is funded by a Medical Research Council UK Clinical Academic Partnership Award. CB-D is in part supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol; the views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. JCM and CM are directly and indirectly supported by the University College London Hospitals and Barts Health NIHR Biomedical Research Centres.

### References

1 O'Mahony C, Jichi F, Pavlou M, et al. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM Risk-SCD). *Eur Heart J* 2014; **35:** 2010–20.

2 Gersh BJ, Maron BJ, Bonow RO, et al. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation* 2011; **124:** e783–831.

3 Elliott PM, Anastasakis A, Borger MA, et al. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *Eur Heart J* 2014; **35:** 2733–79.

4 Cardim N, Galderisi M, Edvardsen T, et al. Role of multimodality cardiac imaging in the management of patients with hypertrophic cardiomyopathy: an expert consensus of the European Association of Cardiovascular Imaging endorsed by the Saudi Heart Association. *Eur Heart J Cardiovasc Imaging* 2015; **16:** 280.

5 Al-Khatib SM, Stevenson WG, Ackerman MJ, et al. 2017 AHA/ACC/HRS guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *Circulation* 2018; **138:** e210–71.

6 Phelan D, Sperry BW, Thavendiranathan P, et al. Comparison of ventricular septal measurements in hypertrophic cardiomyopathy patients who underwent surgical myectomy using multimodality imaging and implications for diagnosis and management. *Am J Cardiol* 2017; **119:** 1656–62.

7 Puntmann VO, Gebker R, Duckett S, et al. Left ventricular chamber dimensions and wall thickness by cardiovascular magnetic resonance: comparison with transthoracic echocardiography. *Eur Heart J Cardiovasc Imaging* 2013; **14:** 240–46.

8 Kramer CM, Barkhausen J, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance (CMR) protocols 2013 update. *J Cardiovasc Magn Reson* 2013; **15:** 91.

9 Suinesiaputra A, Bluemke DA, Cowan BR, et al. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J Cardiovasc Magn Reson* 2015; **17:** 63.

10 Bhuva AN, Bai W, Lau C, et al. A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging* 2019; **12:** e009214.

11 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

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

13 Klinke V, Muzzarelli S, Lauriers N, et al. Quality assessment of cardiovascular magnetic resonance in the setting of the European CMR registry: description and validation of standardized criteria. *J Cardiovasc Magn Reson* 2013; **15:** 55.

14 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017; **39:** 640–51.

15 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *arXiv* 2015; published online May 18. http://arxiv.org/abs/1505.04597.

16 Jones SE, Buchbinder BR, Aharon I. Three-dimensional mapping of cortical thickness using Laplace's equation. *Hum Brain Mapp* 2000; **11:** 12–32.

17 McBride GB. A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report: HAM2005-062; 2005. Hamilton: National Institute of Water & Atmospheric Research, 2005.

18 Jones R, Payne B. Clinical investigation and statistics in laboratory medicine. London: ACB Venture Publications, 1997.

19 Hyslop NP, White WH. Estimating precision using duplicate measurements. *J Air Waste Manag Assoc* 2009; **59:** 1032–39.

20 Forkman J. Estimator and tests for common coefficients of variation in normal distributions. *Commun Stat Theory Methods* 2009; **38:** 233–51.

21 Bunting KV, Steeds RP, Slater LT, Rogers JK, Gkoutos GV, Kotecha D. A practical guide to assess the reproducibility of echocardiographic measurements. *J Am Soc Echocardiogr* 2019; **32:** 1505–15.