Circulation: Cardiovascular Imaging

ORIGINAL ARTICLE

Deep Learning–Based Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction

A Point-of-Care Solution

See Editorial by Fornwalt and Pfeifer

BACKGROUND: We have recently tested an automated machine-learning algorithm that quantifies left ventricular (LV) ejection fraction (EF) from guidelines-recommended apical views. However, in the point-of-care (POC) setting, apical 2-chamber views are often difficult to obtain, limiting the usefulness of this approach. Since most POC physicians often rely on visual assessment of apical 4-chamber and parasternal long-axis views, our algorithm was adapted to use either one of these 3 views or any combination. This study aimed to (1) test the accuracy of these automated estimates; (2) determine whether they could be used to accurately classify LV function.

METHODS: Reference EF was obtained using conventional biplane measurements by experienced echocardiographers. In protocol 1, we used echocardiographic images from 166 clinical examinations. Both automated and reference EF values were used to categorize LV function as hyperdynamic (EF>73%), normal (53%–73%), mildly-to-moderately (30%–52%), or severely reduced (<30%). Additionally, LV function was visually estimated for each view by 10 experienced physicians. Accuracy of the detection of reduced LV function (EF<53%) by the automated classification and physicians' interpretation was assessed against the reference classification. In protocol 2, we tested the new machine-learning algorithm in the POC setting on images acquired by nurses using a portable imaging system.

RESULTS: Protocol 1: the agreement with the reference EF values was good (intraclass correlation, 0.86–0.95), with biases <2%. Machine-learning classification of LV function showed similar accuracy to that by physicians in most views, with only 10% to 15% cases where it was less accurate. Protocol 2: the agreement with the reference values was excellent (intraclass correlation=0.84) with a minimal bias of 2.5±6.4%.

CONCLUSIONS: The new machine-learning algorithm allows accurate automated evaluation of LV function from echocardiographic views commonly used in the POC setting. This approach will enable more POC personnel to accurately assess LV function.

Federico M. Asch, MD
Victor Mor-Avi, PhD
David Rubenson, MD
Steven Goldstein, MD
Muhamed Saric, MD
Issam Mikati, MD
Samuel Surette, BS
Ali Chaudhry, BS, MBA
Nicolas Poilvert, PhD
Ha Hong, PhD
Russ Horowitz, MD
Daniel Park, MD
Jose L. Diaz-Gomez, MD
Brandon Boesch, DO
Sara Nikravan, MD
Rachel B. Liu, MD
Carolyn Philips, MD
James D. Thomas, MD
Randolph P. Martin, MD
Roberto M. Lang, MD

Key Words: algorithm ■ artificial intelligence ■ echocardiography ■ machine learning ■ ventricular function, left

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

CLINICAL PERSPECTIVE

This study showed that the new machine learning algorithm, adapted for echocardiographic views commonly used in the point-of-care setting, allows a fully automated evaluation of left ventricular ejection fraction and, specifically, detection of left ventricular dysfunction with accuracy similar to that of visual interpretation by experienced imaging cardiologists. This technology is likely to enable more health care personnel to confidently perform and interpret bedside cardiac ultrasound examinations, which would address the need driven by the growing availability of handheld ultrasound imaging devices.

Although the limitations of left ventricular (LV) ejection fraction (EF) are well known, it remains the principal clinical echocardiographic measure of LV function. Current echocardiography guidelines[1] emphasize the importance of accurate quantification of LV EF, as multiple indications for therapeutic interventions rely on accurate values of this parameter. Quantitative evaluation of LV EF typically requires measurement of end-systolic and end-diastolic volumes in 2 orthogonal apical views, using traced endocardial boundaries at these 2 phases of the cardiac cycle, followed by model-based calculations (biplane method of disks summation). However, boundary identification is prone to errors due to suboptimal image quality, artifacts, and unusual LV shape in different pathologies, all resulting in considerable interobserver variability.[2–7] To circumvent this problem, we recently tested a newly developed, fully automated machine-learning (ML) algorithm that estimates LV EF without tracing the endocardial boundaries and measuring LV volumes.[8] This approach is similar to a human expert reader visually estimating the degree of ventricular contraction in the apical 2- and 4-chamber (AP2, AP4) views, based on years of experience acquired through the interpretation of thousands of exams, and translating this information into a quantitative LV EF value.

However, in the point-of-care (POC) setting, acquisition of the AP2 view by nonechocardiographers is challenging, and as a result, most POC clinicians usually rely in the evaluation of LV function on visual assessment of easier-to-obtain views, such as parasternal long-axis (PLAX) and/or AP4. Accordingly, to make the new ML algorithm useful in the POC setting, the previously tested software was modified to quantify LV EF from any one of the above 3 long-axis views, or any combination of these views available in an individual patient. The primary goal of this study was to test the accuracy of these fully automated measurements in unselected patients referred for a clinically indicated echocardiogram. The secondary goal was to determine whether this approach would provide added diagnostic accuracy in the POC setting.

METHODS

Data and materials used in this study will not be made publicly available.

Study Design

The new ML algorithm (Caption Health, Inc, San Francisco, CA) was tested in 2 separate protocols. Protocol 1 was designed to initially test the performance of the algorithm on images acquired by cardiac sonographers using systems typically found in echocardiography laboratories. First, automated EF measurements were compared with reference values provided by expert readers using conventional methodology. Then, we tested the ability to categorize LV function based on these automated measurements using commonly clinically utilized classification of LV function by discrete grades. Although this methodology is highly subjective, studies have shown that when performed by expert readers, it may be relatively accurate compared with actual measurements.[2–4,6,7,9,10] To place this approach into practical perspective, we compared the accuracy of the automated ML-based classification of LV function to that based on visual assessment by 2 groups of physicians, including both trained imaging cardiologists and POC clinicians.

Protocol 2 was designed to test the new ML algorithm in the POC setting. In this protocol, the input was images prospectively acquired by POC nurses using a portable imaging system, which combines real-time ML-based guidance for image acquisition from anatomically correct views[11] with automated ML-based quantification of LVEF. These images were automatically analyzed to obtain EF values, which were compared, as in Protocol 1, to reference values provided by expert readers using conventional methodology. Accuracy metrics obtained in this protocol were compared with those noted in Protocol 1 to assess the extent to which the use of POC images obtained by less skilled users without high-end equipment would impact the performance of the ML algorithm.

ML Approach

Our ML algorithm was developed to estimate LV EF without measuring LV end-systolic and end-diastolic volumes.[8] Briefly, this alternative approach assumes that the ventricle contracts throughout systole simultaneously along its long axis and in the radial direction, so that its corresponding dimensions L and R change over time according to 2 dimensionless contraction coefficients C_L and C_R (Figure 1). Using these coefficients, LV volume at end-systole can be described by:

$$V(ES) = V(ED) \cdot C_L \cdot C_R$$

where V(ED) is the volume at end-diastole. By definition of LV EF as the difference between V(ED) and V(ES) normalized by the former, it can be expressed in terms of the above 2 contraction coefficients as follows:

$$EF = \left[V(ED) - V(ES)\right] / V(ED) = 1 - V(ES)/V(ED) = 1 - C_L \cdot C_R$$

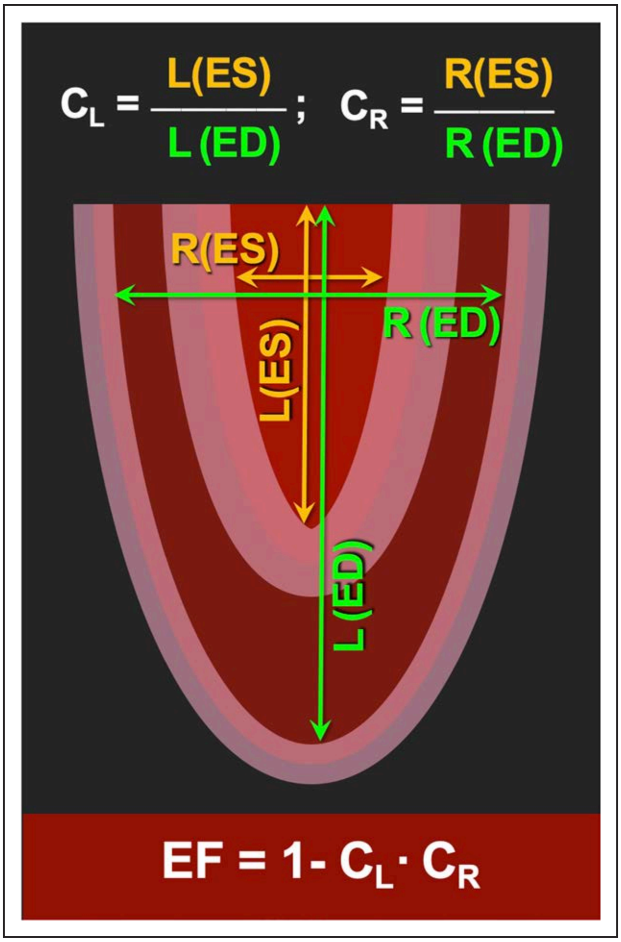

Figure 1. Schematic depiction of the principle of the machine-learning (ML) algorithm for automated quantification of ejection fraction (EF) without measuring left ventricular (LV) volumes.
Contraction coefficients in the longitudinal and radial directions, CL and CR, defined as the ratios between ventricular dimensions, L and R, at end-systole (ES) and end-diastole (ED) are estimated and used to calculate EF. This is based on the assumption that LV volume changes from ED to ES according to the changes in these dimensions, that is, V(ES)=V(ED)·CL·CR. For example, if during systole, the ventricle shortens by 14%, CL would be 0.86, and if at the same time its radial dimension shortens by 30%, corresponding to CR value of 0.70, this would result in an EF of 40%: EF=1–[0.86, 0.70]=1–0.60=0.40.

Which allows the calculation of EF without measuring volumes from the estimated contraction coefficients in the longitudinal and radial direction in the apical views. In the PLAX view, where the longitudinal contraction is difficult to estimate without the apex being fully visualized, CL is estimated from the contraction of the basal portion of the ventricle.

Our ML algorithm was trained the computer to estimate the above 2 contraction coefficients, C_L and C_R. Briefly, we used a deep learning technique, which does not use any sort of explicit tracking methodology, but instead lets the neural network decide from the data itself what the best approach to handle the data would be. In other words, the algorithm was not guided by the developers as to what should be detected or tracked throughout the cardiac cycle. Instead, the algorithm was allowed to derive from thousands of images the features and visual patterns necessary to estimate EF in agreement with the reference values obtained by human readers using conventional methodology. The neural network was constrained to report the amplitude of change in ventricular dimensions, roughly the equivalent of the above contraction coefficients. Importantly, the neural network had the total freedom in choosing to track relative sizes/dimensions of physiological features and/or speckle patterns.

The algorithm developed on the basis of the above described principle (Caption Health, Brisbane, CA) was implemented in Python and trained using Keras (https://keras.io/) with a Tensorflow (https://www.tensorflow.org/) backend to train and deploy the Neural Networks. The training was performed on a database of >50 000 echocardiographic studies. Training included the use of multiple AP2, AP4, and PLAX views available as part of each individual exam and LV EF values measured over the years by clinicians interpreting these studies using conventional methodology (biplane Simpson technique. Following this training, the algorithm was designed to provide fully automated estimates of LV EF on any combination of the above 3 views.

Protocol 1

In this protocol, we used echocardiographic images of 166 patients (age, 20–90 years; 97 males, 69 females; median body mass index, 26.7 kg/m^2, range, 16.0–48.4 kg/m^2), including inpatients and outpatients, who underwent clinically indicated echocardiographic examinations at one of the 3 participating institutions (Duke University Medical Center, Northwestern Memorial Hospital, and Minneapolis Heart Institute). These studies were selected to equally represent a wide range of body mass indices and LV function, reflected by a wide range of LV EF. Image quality was not used to exclude patients. No additional criteria were used to select images, to test the software in a cohort reflective of the general population of patients referred for an echocardiogram. Images were acquired using a random mix of available equipment, including Philips, GE, and Siemens imaging systems (Philips/CX50 4 [2.4%], Acuson/SEQUOIA 27 [16.3%], Philips/EPIQ 7C 1 [0.6%], GE/Vivid i 22 [13.3%], GE/Vivid E9 3 [1.8%], GE/Vivid7 9 [5.4%], Siemens/ACUSON SC2000 8 [4.8%], Philips/iE33 92 [55.4%]). This protocol was approved by the Institutional Review Boards of the 3 institutions with a waiver of consent.

To establish a reference standard for LV EF, images of each patient were independently analyzed by 3 experienced sonographers who traced the endocardial boundaries in the AP4 and AP2 views and calculated LV EF using the guidelines-recommended biplane method of disks. These measurements were then reviewed and, if necessary, corrected by 3 experienced, board-certified imaging cardiologists. These biplane measurement sets were averaged to provide a unique, single reference EF value for each patient.

The validation of the ML algorithm in this protocol included 3 parts. First, ML-generated EF measurements were compared against the reference EF values using intraclass correlation and Bland-Altman analysis of biases and limits of agreement. In addition, as an alternative to Bland-Altman bias, which can be zero in a large sample in the presence of wide limits of agreement, inter-technique agreement was also assessed by calculating mean absolute difference (MAD) by averaging absolute differences between the ML estimates and the 3 experts' individual measurements. Furthermore, inter-technique agreement was assessed by determining the number of outliers, namely cases in which EF discrepancy was >10% and separately >15%. These comparisons against the above reference standard were performed separately for the automated single-view analyses, including AP4, AP2, and PLAX views, as well as for all possible combinations of 2 views (AP4+AP2; PLAX+AP4; PLAX+AP2) and jointly for all 3 views.

Second, both the automated and the reference EF values were used to classify LV function in each case as: hyperdynamic (EF>73%), normal (EF between 53% and 73%), mildly-to-moderately reduced (EF from 30% to 52%), or severely reduced (EF<30%). The agreement between the 2 classifications was also tested for single views, as well as for all the aforementioned view combinations, by calculating the percentage of cases where the automated classification and the expert classification provided the same result out of the total number of cases.

Finally, every image sequence was reviewed by 10 physicians, including 3 experienced imaging cardiologists (not the same ones who participated in establishing the reference values) and 7 physicians experienced in POC cardiac ultrasound (emergency medicine and critical care medicine). These 10 physicians independently reviewed each image sequence, visually estimated EF, and graded LV function accordingly for each individual view, using the same 4 categories described above. Each of these grades was then compared with the reference grades, resulting in a total of 1660 comparisons for each view or combination of views, including 498 by the cardiologists and 1162 by the POC physicians. Agreement levels between the physicians' visual classifications and the reference classification were then compared with those between the automated ML classification and the same reference. This was completed by separately evaluating the findings of the 3 cardiologists and those of the 7 POC physicians. These comparisons included a detailed analysis of the numbers (and percentages) of comparisons in which the ML-generated classification was equally accurate, more accurate, and less accurate than the physicians' visual classifications.

Protocol 2

In this protocol, we prospectively studied 67 patients (see Table 1 for demographics and basic characteristics) who, in addition to their clinically indicated cardiac ultrasound examination, underwent imaging by one of 8 nurses. Imaging was performed in the echocardiography laboratory or the inpatient wards. The nurses had no prior experience with ultrasound imaging and underwent a 1-hour didactic training session and 12 practice scans, during which they received hands-on instruction on the use of the ML software that provides real-time prescriptive guidance to optimize transducer position and orientation for acquisition of 3 standard echocardiographic views (A4C, A2C, and PLAX), obtained using a portable ultrasound imaging system (Terason uSmart 3200t Plus).[11] This system automatically captures the image, when it is determined to be anatomically correct. Alternatively, if the user cannot achieve this, he/she can prompt the software to use the best view

seen throughout the scan.

All automatically saved images showing anatomically correct views were used for automated analysis by the new ML algorithm, which provided EF values for each individual view, as well as for the combination of all views available in each patient. In addition, the AP4 view and, when available, AP2 view were assessed using conventional methodology by 3 expert readers, whose EF values were averaged and used as a reference EF value for each patient. Agreement with the reference was quantified using intraclass correlation and Bland-Altman analyses. This protocol was approved by the Institutional Review Board and each patient signed informed consent.

Table 1. Demographics, Basic Characteristics, and Clinical Findings of the Patients in Protocol 2

Patient characteristics	
N	
Sex (male, female)	35, 32
Age, y	61±17
Height, cm	169±9
Weight, kg	80±16
BMI, kg/m²	28±6
Race, %	
White participants	76.1
Black participants	14.9
Asian	1.5
Other	7.5
Valve abnormalities, %	
Mitral	52
Aortic	43
Tricuspid	54
Cardiac abnormality, %	88
LV hypertrophy, %	48
Pericardial effusion, %	4.5
Patent foramen ovale, %	1.5

BMI indicates body mass index; and LV, left ventricular.

Statistical Analysis

Data were expressed as mean values±SDs or median (range), when appropriate. The automated EF estimates were compared with the reference standard using intraclass correlation (ICC), Bland-Altman analyses, and MAD. Finally, we calculated the sensitivity, specificity, negative and positive predictive values (NPV and PPV) of the detection of reduced LV function, reflected by EF<53%. All analyses, including basic statistics, were performed using the Python Scipy.Stats package (Python Software Foundation, Beaverton, OR).

RESULTS

Protocol 1

Reference values of LV EF in the study group varied from 15% to 77% with a median value of 46%, and were uniformly distributed over the entire range. The algorithm was able to analyze at least one view in every one of the 166 patients. The time to obtain an automated EF value for a given view was in the order of magnitude of 1 to 5 seconds on a standard personal computer. The results of the intraclass correlation and Bland-Altman analyses are shown in Figures 2 and 3, which reflect overall very good agreement between the automated EF measurements and the reference values, with high correlations (ICC range, 0.86–0.95), minimal biases (<2%), and reasonable limits of agreements.

Table 2 shows the calculated MAD values between the automated EF measurements and the corresponding reference values for each view or combination of views (second column). Differences were larger for single views compared with the measurements based on view combinations, although there was a certain level

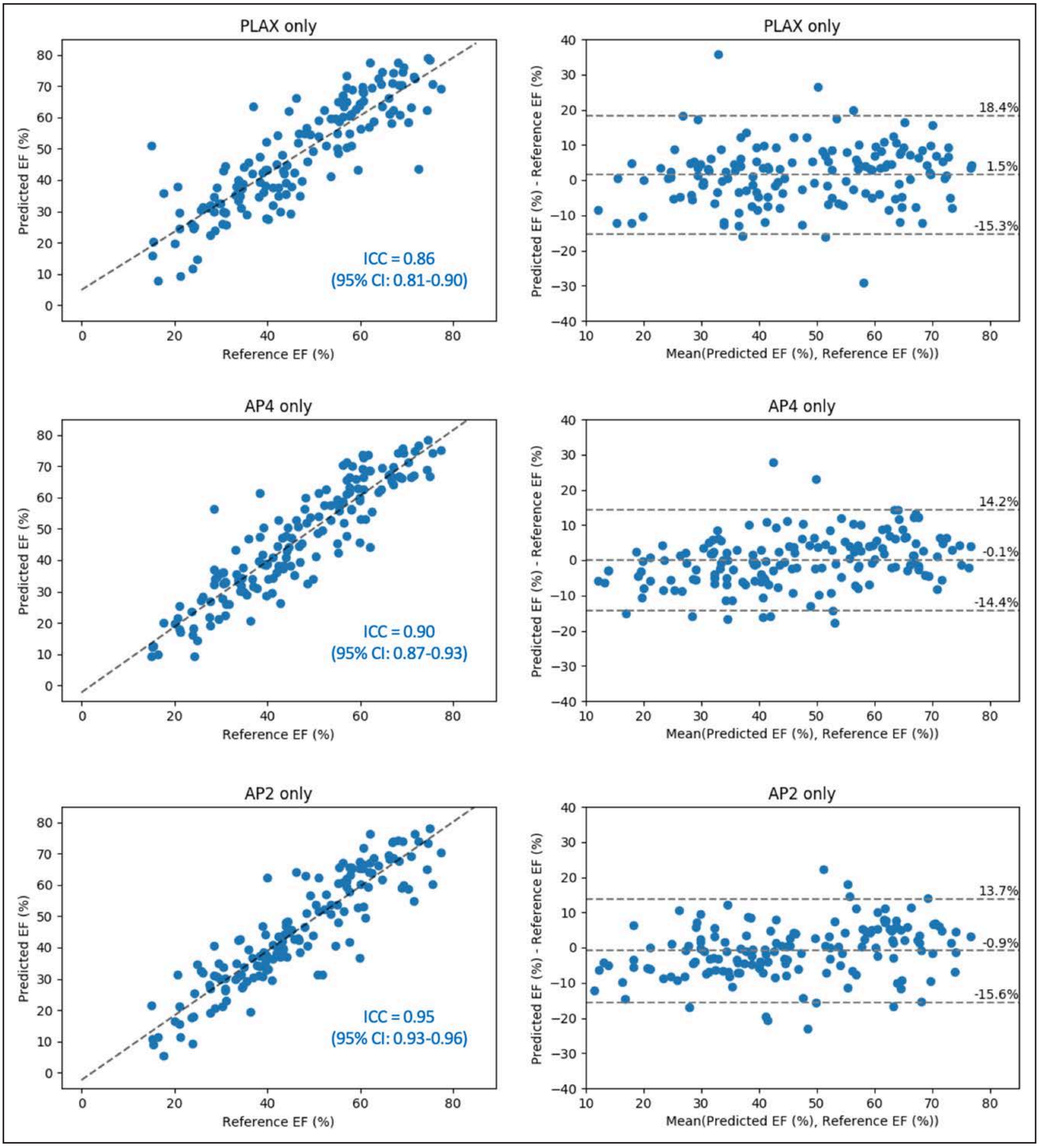

Figure 2. Agreement between the machine learning based automated ejection fraction (EF) measurements and reference values: intraclass correlation (ICC, left) and Bland-Altman analysis (right).
Data shown for the 3 single views. PLAX indicates parasternal long-axis.

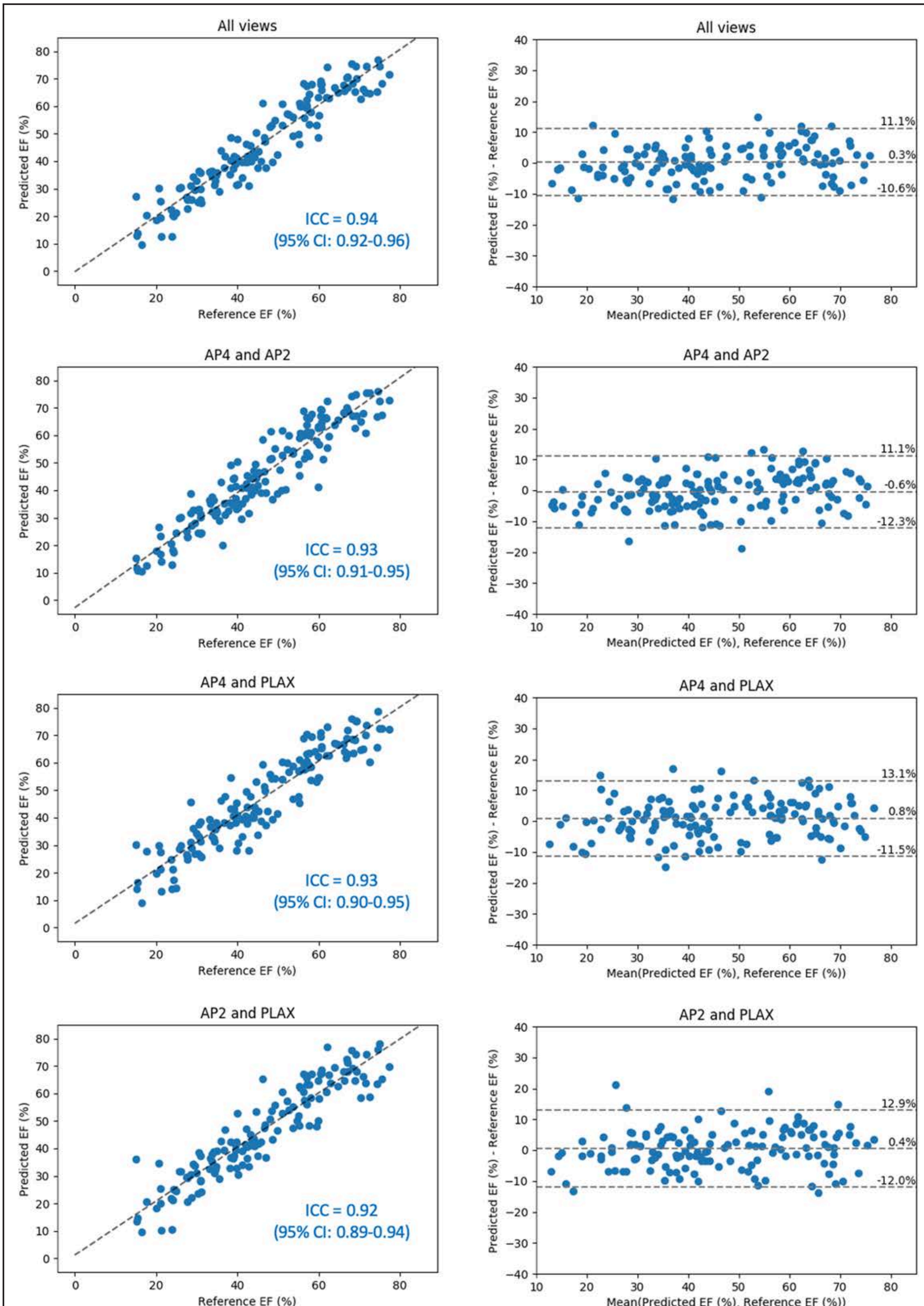

Figure 3. Agreement between the machine learning based automated ejection fraction (EF) measurements and reference values: intraclass correlation (ICC, left) and Bland-Altman analysis (right).
Data shown for the 4 possible combinations of 2 and all 3 views. PLAX indicates parasternal long-axis.

of overlap between the CIs in some of the comparisons. Importantly, all MAD values were <7%, including those based on a single PLAX view, for which the largest difference was noted. This was similar to MAD between the experts' individual biplane measurements, which was 9%.

Finally, analysis of outliers showed that the percentage of cases in which EF error was >15% was 0 for the combination of all 3 views, between 1% and 2% for any combination of 2 views, and higher but below 8% for single-view estimates, with the largest noted in the PLAX view at 7.5% (Table 2, right column). We found no clinical factors to be significantly associated with less accurate EF estimates, which were clearly driven by suboptimal image quality.

Based on the reference measurements, LV function was classified as hyperdynamic in 5 patients, normal in 61, mildly-to-moderately reduced in 73 and severely reduced in 27 patients. The overall accuracy numbers for the automated, ML-based classification of LV function versus the physicians' classification against the same reference standard are shown in Table 3. Interestingly, cardiologists' classification of LV function resulted in higher accuracy metrics than that of POC physicians using the apical views, while the POC physicians' classifications suggested higher accuracy than those of the cardiologists' when using the PLAX view. The automated ML classifications resulted in higher accuracy metrics than both groups of physicians for both apical views, while in the PLAX view, the accuracy of the ML classifications surpassed that of the cardiologists' and was similar to that of the POC physicians.

In terms of the ability to correctly identify patients with reduced LV function (EF<53%) from single views, the detailed comparison of performance metrics showed that the automated analysis of single views resulted in high sensitivity, specificity, NPV, and PPV, reflecting a comparable to or better diagnostic performance than that of the 2 groups of physicians for most metrics (Table 4).

Table 5 shows a summary of the comparisons between the accuracy of the automated classification versus the physicians' classification of LV function against the same reference standard by numbers (and percentages) of cases, in which the ML-generated classification was more accurate, equally accurate, and less accurate than the physicians' visual classifications. Importantly, for all 3 views, the percentages of ML classification less accurate than those by either group of physicians was consistently in the low teens (>10% but <15%), while ML classification was more accurate than the physicians' classification in ≈20% of the cases, indicating overall not only very good, but, in fact, better performance.

Protocol 2

The nurses were able to acquire at least 1 view suitable for analysis, resulting in an automated EF value in every one of the 67 patients. Of these patients, AP4 views of sufficient image quality to produce an automated EF estimate were available in 59, AP2 views in 38, and PLAX views in 37 patients. Overall, the agreement between the automated EF measurements and the reference values was excellent, as reflected by an ICC-value of 0.84 (CI, 0.75–0.90), a minimal bias of 2.5±6.4% (limits of agreement: −10.4% to 15.3%), and MAD of 5.4% (Figure 4). Single-view analysis showed that EF values derived from the PLAX view were slightly less accurate than those obtained from the 2 apical views with the lowest correlation and largest bias noted in the PLAX view (Table 6). When compared with the accuracy of the ML algorithm with images acquired by cardiac sonographers using high-end equipment in Protocol 1, the agreement with the reference was lower, as shown by the nonoverlapping CIs: ICC=0.84 [95% CI, 0.75–0.90] versus ICC=0.94 [95% CI, 0.92–0.96] and the bias larger (2.5% versus 0.3%) with slightly wider limits of agreement.

DISCUSSION

In this study, we tested the accuracy of a novel fully automated ML algorithm for the quantification of LV EF from long-axis views, including the PLAX view, which is commonly used in the POC setting. The testing included single views analyzed individually and using all possible combinations, to reflect different scenarios encountered in daily clinical practice including POC ultrasound imaging. These automated measurements were compared with reference values obtained by conventional biplane measurements performed and verified by expert echocardiographers. Our key findings in Protocol 1 were (1) the automated ML-based measurements are reasonably accurate, including the PLAX view, in which the entire LV apex is frequently not visualized, (2) these automated measurements allowed accurate classification of LV systolic function into one of 4 categories, as well as detection of LV dysfunction with accuracy similar to or even better than that of experienced physicians, including both cardiologists and POC ultrasound users. Protocol 2 showed that automated ML-based EF measurements are almost as accurate, when using images acquired by less skilled POC users, such as nurses, using a typical portable imaging system with the built-in real-time prescriptive guidance software.

The use of echocardiographic imaging for quick bedside assessment of LV function in the POC setting has been steadily increasing over the past decade with the widespread availability of affordable and user-friendly handheld ultrasound imaging devices, which lend themselves to a quick bedside assessment of LV function.[12] However, training of POC personnel in both image acquisition and interpretation of fundamental components such as LV function varies widely. In addition, environmental challenges and patient factors prevalent in noncardiologist clinical areas result in limited image

Table 2. Agreement Between the Automated EF Measurements and the Reference Values for Each View or Combination of Views: (MAD, Second Column; Values Are Median With the Corresponding Ranges in Brackets); Analysis of Outliers Showing Percentages of Cases in Which EF Error Was >10% (Third column) and >15% (Right Column)

Views	EF MAD (%) (95% CI)	Cases with error >10% (95% CI)	Cases with error >15% (95% CI)
PLAX, AP4, and AP2	4.46 (3.91–5.02)	6.7 (3.0–11.1)	0.0 (0.0–0.0)
AP4 and AP2	4.89 (4.34–5.45)	12.5 (7.2–17.8)	1.3 (0.0–3.3)
AP4 and PLAX	5.10 (4.48–5.71)	12.5 (7.6–18.1)	1.4 (0.0–3.5)
AP2 and PLAX	5.05 (4.40–5.70)	10.9 (5.8–16.1)	1.5 (0.0–3.6)
AP4	5.64 (4.92–6.35)	16.7 (11.1–22.8)	4.9 (1.9–8.6)
AP2	5.89 (5.15–6.63)	14.8 (9.7–20.6)	5.8 (2.6–9.7)
PLAX	6.76 (5.87–7.66)	19.7 (13.6–26.5)	7.5 (3.4–12.2)

See text for details. EF indicates ejection fraction; MAD, mean absolute difference; and PLAX, parasternal long-axis.

Table 4. Sensitivity, Specificity, NPV, and PPV of the Single-View Detection of Reduced LV Function, Reflected by EF<53%

View		Automated classification	Cardiologists' classification	POC physicians' classification
AP4	Sensitivity, %	91.0	84.7	72.4
	Specificity, %	83.3	88.4	87.3
	PPV, %	92.9	94.4	94.8
	NPV, %	85.9	79.9	70.1
AP2	Sensitivity, %	89.0	80.3	68.3
	Specificity, %	75.8	89.1	91.5
	PPV, %	90.8	95.3	96.2
	NPV, %	87.7	80.3	67.8
PLAX	Sensitivity, %	77.0	79.6	78.6
	Specificity, %	78.8	85.1	86.4
	PPV, %	91.7	89.7	94.6
	NPV, %	82.5	82.6	79.9

Values for the automated ML detection, as well as for the 2 groups of physicians, are shown separately for each view. See text for details. EF indicates ejection fraction; LV, left ventricular; NPV, negative predictive value; PLAX, parasternal long-axis; POC, point-of-care; and PPV, positive predictive value.

quality with potentially subsequent erroneous assessments. Accordingly, computer-assisted solutions designed to guide nonechocardiographers in these tasks have been increasingly sought after. One such potential solution is the deep learning algorithm recently validated by our group.[8] In the current study, we tested an advanced version of this algorithm that was specifically adapted to expand its utility to the POC environment.

As the number of POC clinicians performing bedside ultrasound for cardiac function in the hospital-wide practice (including emergency department, intensive care units, postanesthesia care units, step-down units, outpatient clinics) grows, the need for more accurate determination of LVEF increases in parallel. The ML software used in this study was designed to address this need. ML technology has been rapidly proving its usefulness in cardiac imaging, where it lends itself to an increasing variety of automated measurements that traditionally relied on extensive user input.[13,14] It is conceivable that ML tools such as the one validated in this study will become widely used in the POC setting and allow accurate echocardiographic evaluation of LV function by increasing numbers of health care providers, particularly when combined with evolving technologies aimed to assist in image acquisition. These tools are especially important in the current era of the rampant COVID-19 pandemic, when bedside assessment is an urgent clinical issue complicated by the need for strict infection control.

Table 3. Overall Accuracy of the Automated, Single-View ML-Based Classification of LV Function vs the Physicians' Classification Against the Same Reference Standard

View	Accuracy of automated classification	Accuracy of cardiologists' classification	Accuracy of POC physicians' classification
AP4	75.3%	72.4%	63.9%
AP2	71.6%	68.9%	62.4%
PLAX	66.7%	61.7%	67.7%

Physicians' data shown by specialty: 3 cardiologists and separately 7 POC physicians. Numbers represent percentages of comparisons resulting in identical classifications as the reference classification. LV indicates left ventricular; ML, machine-learning; PLAX, parasternal long-axis; and POC, point-of-care.

In this setting, a key issue addressed in our study was the ability of a new ML algorithm to estimate LV EF from the PLAX view, which is often easier to obtain, routinely used in the POC setting but is not recommended by the guidelines for the assessment of LV function because of the incomplete visualization of the LV apex. Interestingly, we found that POC physicians' classification of LV function from this view alone was only slightly less accurate than the visual assessment from the usual apical views performed by expert imaging cardiologists (Table 2). Importantly, in the context of this study, in all 3 views, the automated classification by the ML algorithm was as accurate if not better than that obtained by both groups of physicians. These findings encourage additional optimization of POC echocardiographic training to assess LV EF from both the PLAX and apical views to optimize clinical care and underscore the effectiveness of the ML algorithm.

Furthermore, we found that the combined assessment of LV function from > 1 view further improved the accuracy of the automated measurements. When combined with the deterministic nature of the fully automated process, these are considerable strengths of the ML approach. This approach is feasible in the majority of patients and provides additional strength to the novel algorithm compared with visual interpretation, which is frequently difficult for inexperienced readers and even for expert cardiologists.

The main limitation of Protocol 1 was that testing was performed on images obtained by experienced cardiac

Table 5. Numbers (and Percentages) of the 1660 Comparisons Made on 166 Sets of Images by 3 Cardiologists and 7 POC Physicians, in Which the ML-Generated Classification Was More Accurate, Equally Accurate, and Less Accurate Than the Physicians' Visual Classifications

	Cardiologists' classification			POC physicians' classification		
	AP4, n (%)	AP2, n (%)	PLAX, n (%)	AP4, n (%)	AP2, n (%)	PLAX, n (%)
ML better	82 (16.5%)	71 (14.3%)	88 (17.7%)	258 (22.2%)	252 (21.7%)	154 (13.3%)
ML equal	327 (65.7%)	311 (62.4%)	275 (55.2%)	692 (59.6%)	650 (55.9%)	658 (56.6%)
ML worse	69 (13.9%)	60 (12.0%)	65 (13.1%)	133 (11.4%)	142 (12.2%)	164 (14.1%)
No assessment	20 (4.0%)	56 (11.2%)	70 (14.1%)	79 (6.8%)	118 (10.2%)	186 (16.0%)
Total	498 (100%)			1162 (100%)		

The No assessment line represents the cases in which physicians were unable to classify LV function due to suboptimal image quality. ML indicates machine-learning; PLAX, parasternal long-axis; and POC, point-of-care.

sonographers, which are likely to be of higher quality than those obtained by less skilled imagers. Also, in Protocol 1, all images were acquired on standard commercial ultrasound imaging systems, which provide better quality images than devices commonly used in the POC setting. Accordingly, Protocol 2 was designed to determine the feasibility rates and accuracy metrics when testing the ML algorithm on images obtained in an actual POC setting using a portable imaging device. It is not surprising that when anatomically correct images acquired by any user irrespective of skill level with the aid of the real-time prescriptive guidance system, the automated analysis provides accurate EF values. These 2 protocols together represent an important first step of validation of this potentially clinically useful technology that incorporates the PLAX view in the automated analysis of LV EF.

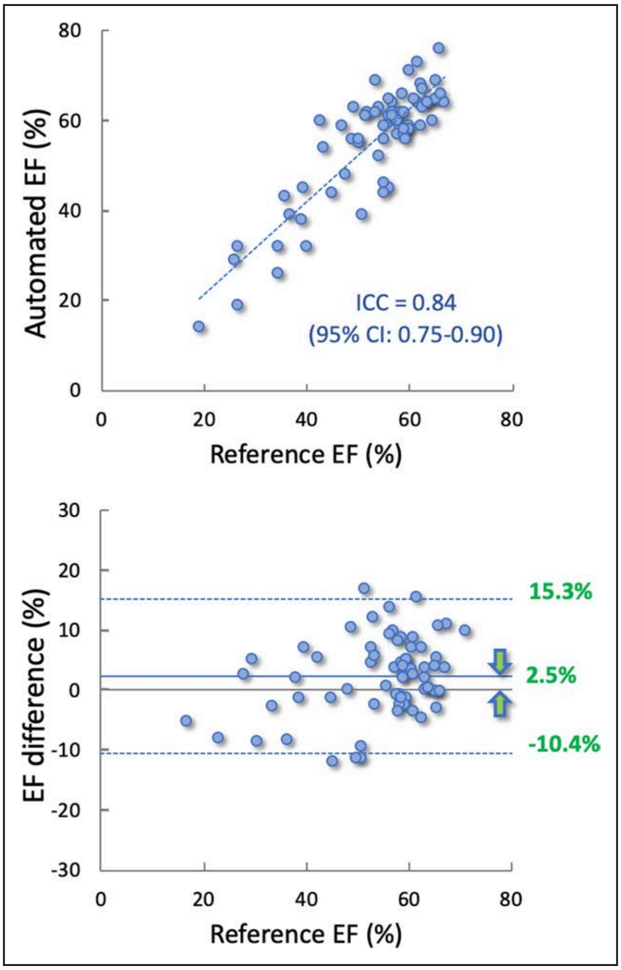

Figure 4. Agreement between the machine learning based automated ejection fraction (EF) measurements and reference values obtained by point-of-care clinicians using a portable imaging system: intraclass correlation (ICC, top) and Bland-Altman analysis (bottom).
Data shown for the combination of all available views in each patient.

Limitations

One limitation of the software tested in this study is that it does not assess other parameters of LV function, beyond EF, and therefore this study cannot answer the question whether automated measurement of additional functional parameters would further improve the evaluation of patients undergoing POC ultrasound.

One might see as a limitation the fact that cardiac magnetic resonance, which is frequently referred to as the gold standard for cardiac chamber quantification,[15] was not used as a reference for comparisons in our study. However, this choice was made consciously during the design of the study, and was directly related to our aim of testing the novel ML algorithm as a potential substitute for the prevailing methodology currently used in the POC setting. We felt that the best way to achieve this goal was to compare both techniques side-by-side against a strong echocardiography-based reference, created by averaging EF measurements obtained by 3 sonographers using the gold-standard biplane method of disks and then verified by expert cardiologists. Comparisons against this

Table 6. Results of ICC and Bland-Altman Analyses of Agreement Between Automated EF Measurements by Nurses and the Reference EF Values for Each View or Combination of All Available Views

	All available views	AP4	AP2	PLAX
ICC	0.84	0.83	0.84	0.74
95% CI	0.75 to 0.90	0.73 to 0.90	0.71 to 0.91	0.55 to 0.86
Bias, %	2.5	1.0	2.9	4.2
LOA, %	−10.4 to 15.3	−14.7 to 16.7	−12.2 to 17.9	−9.4 to 17.7

EF indicates ejection fraction; ICC, intraclass correlation; LOA, level of agreement; and PLAX, parasternal long-axis reference showed very good agreement, indicating that the automated approach is at least as good as human readers, with the advantages of high feasibility and fully automated nature resulting in perfect reproducibility.

CONCLUSIONS

In summary, this study showed that the new ML algorithm, adapted for echocardiographic views commonly used in the POC setting, allows automated evaluation of LV EF and, specifically, detection of LV dysfunction with accuracy similar to that of visual interpretation by experienced imaging cardiologists. This technology is likely to enable more health care personnel who are developing their competence in POC to confidently perform and interpret bedside cardiac ultrasound examinations, which would address the need driven by the growing availability of handheld imaging devices.

ARTICLE INFORMATION

Received December 10, 2020; accepted April 6, 2021.

Correspondence

Roberto M. Lang, MD, 5758 S Maryland Ave, MC 9067, Room 5509, Chicago, IL 60637. Email rlang@medicine.bsd.uchicago.edu

Affiliations

MedStar Health Research Institute, Washington, DC (F.M.A.). University of Chicago, IL (V.M.-A., R.M.L.). Scripps Clinic and Prebys Cardiovascular Institute, La Jolla, CA (D.R.). MedStar Washington Hospital Center, DC (S.G., C.P.). New York University Langone Health (M.S.). Feinberg School of Medicine, Northwestern University, Chicago, IL (I.M., R.H., J.D.T.). Caption Health Inc, San Francisco, CA (S.S., A.C., N.P., H.H., R.P.M.). University of North Carolina Medical Center (D.P). Baylor St. Luke's Medical Center, Houston, TX (J.L.D.-G.). Highland Hospital, Oakland, CA (B.B.). University of Washington Medical Center, Seattle (S.N.). Yale School of Medicine, New Haven, CT (R.B.L.). Emory University Medical Center, Atlanta, GA (R.P.M.).

Acknowledgments

We wish to thank Dr Richard Bae at Minneapolis Heart Institute, and Dr James G. Jollis at Duke University Medical Center for providing the datasets used in this study, as well as Ms Birgitta Cheng and Dr Tova Brown for her help with data analysis.

Sources of Funding

This work has been partially supported by Caption Health.

Disclosures

In addition to employment relationships between several authors (S. Surette, A. Chaudhry, N. Poilvert, Dr Hong, Dr Martin) and Caption Health, Inc, Mor-Avi, Thomas, and Lang served as paid consultants to this company for work related to this article. Dr Asch received institutional research grants or unpaid research collaborations with Caption Health, TOMTEC, Butterfly, and Ultromics. Dr Lang also received institutional research grants from Philips Medical Systems and TOMTEC, all unrelated to this study. The other authors report no conflicts.

REFERENCES

1. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr*. 2015;28:1–39.e14. doi: 10.1016/j.echo.2014.10.003
2. Rich S, Sheikh A, Gallastegui J, Kondos GT, Mason T, Lam W. Determination of left ventricular ejection fraction by visual estimation during real-time two-dimensional echocardiography. *Am Heart J*. 1982;104:603–606. doi: 10.1016/0002-8703(82)90233-2
3. Hope MD, de la Pena E, Yang PC, Liang DH, McConnell MV, Rosenthal DN. A visual approach for the accurate determination of echocardiographic left ventricular ejection fraction by medical students. *J Am Soc Echocardiogr*. 2003;16:824–831. doi: 10.1067/S0894-7317(03)00400-0
4. Mele D, Campana M, Sclavo M, Seveso G, Aschieri D, Nesta F, D'Aiello I, Ferrari R, Levine RA. Impact of tissue harmonic imaging in patients with distorted left ventricles: improvement in accuracy and reproducibility of visual, manual and automated echocardiographic assessment of left ventricular ejection fraction. *Eur J Echocardiogr*. 2003;4:59–67. doi: 10.1053/euje.2002.0619
5. Mor-Avi V, Jenkins C, Kühl HP, Nesser HJ, Marwick T, Franke A, Ebner C, Freed BH, Steringer-Mascherbauer R, Pollard H, et al. Real-time 3-dimensional echocardiographic quantification of left ventricular volumes: multicenter study for validation with magnetic resonance imaging and investigation of sources of error. *JACC Cardiovasc Imaging*. 2008;1:413–423. doi: 10.1016/j.jcmg.2008.02.009
6. Thavendiranathan P, Popović ZB, Flamm SD, Dahiya A, Grimm RA, Marwick TH. Improved interobserver variability and accuracy of echocardiographic visual left ventricular ejection fraction assessment through a self-directed learning program using cardiac magnetic resonance images. *J Am Soc Echocardiogr*. 2013;26:1267–1273. doi: 10.1016/j.echo.2013.07.017
7. Kusunose K, Shibayama K, Iwano H, Izumo M, Kagiyama N, Kurosawa K, Mihara H, Oe H, Onishi T, Onishi T, et al; JAYEF Investigators. Reduced variability of visual left ventricular ejection fraction assessment with reference images: the Japanese Association of Young Echocardiography Fellows multicenter study. *J Cardiol*. 2018;72:74–80. doi: 10.1016/j.jjcc.2018.01.007
8. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, Romano N, Hong H, Mor-Avi V, Martin RP, et al. Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert. *Circ Cardiovasc Imaging*. 2019;12:e009303. doi: 10.1161/CIRCIMAGING.119.009303
9. Moore CL, Rose GA, Tayal VS, Sullivan DM, Arrowood JA, Kline JA. Determination of left ventricular function by emergency physician echocardiography of hypotensive patients. *Acad Emerg Med*. 2002;9:186–193. doi: 10.1111/j.1553-2712.2002.tb00242.x
10. Charron C, Templier F, Goddet NS, Baer M, Vieillard-Baron A; Group of Investigators of SAMU 92. Difficulties encountered by physicians in interpreting focused echocardiography using a pocket ultrasound machine in prehospital emergencies. *Eur J Emerg Med*. 2015;22:17–22. doi: 10.1097/MEJ.0000000000000153
11. Narang A, Bae R, Hong H, Thomas Y, Surette S, Cadieu C, Chaudhry A, Martin RP, McCarthy PM, Rubenson DS, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol.* 2021. doi: 10.1001/jamacardio.2021.0185
12. Prada G, Fritz AV, Restrepo-Holguín M, Guru PK, Díaz-Gómez JL. Focused cardiac ultrasonography for left ventricular systolic function. *N Engl J Med*. 2019;381:e36. doi: 10.1056/NEJMvcm1802841
13. Liu Y, Gopalakrishnan V. An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data (Basel)* 2017;2:8. doi: 10.3390/data2010008
14. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. 2018;138:1623–1635. doi: 10.1161/CIRCULATIONAHA.118.034338
15. Kramer CM, Barkhausen J, Flamm SD, Kim RJ, Nagel E; Society for Cardiovascular Magnetic Resonance Board of Trustees Task Force on Standardized Protocols. Standardized cardiovascular magnetic resonance imaging (CMR) protocols, society for cardiovascular magnetic resonance: board of trustees task force on standardized protocols. *J Cardiovasc Magn Reson*. 2008;10:35. doi: 10.1186/1532-429X-10-35