Human versus Artificial Intelligence–Based Echocardiographic Analysis as a Predictor of Outcomes: An Analysis from the World Alliance Societies of Echocardiography COVID Study

Federico M. Asch, MD, FASE, Tine Descamps, PhD, Rizwan Sarwar, MRCP, Ilya Karagodin, MD, Cristiane Carvalho Singulane, MD, Mingxing Xie, MD, PhD, Edwin S. Tucay, MD, Ana C. Tude Rodrigues, MD, Zuilma Y. Vasquez-Ortiz, MD, PhD, Mark J. Monaghan, MD, Bayardo A. Ordonez Salazar, MD, Laurie Soulat-Dufour, MD, Azin Alizadehasl, MD, Atoosa Mostafavi, MD, Antonella Moreo, MD, Rodolfo Citro, MD, Akhil Narang, MD, Chun Wu, MD, PhD, Karima Addetia, MD, Ross Upton, PhD, Gary M. Woodward, PhD, and Roberto M. Lang, MD, on Behalf of the WASE-COVID Investigators, *Washington, District of Columbia; Oxford and London, United Kingdom; Chicago, Illinois; Wuhan, China; Quezon City, Philippines; São Paulo, Brazil; Mexico City, Mexico; Paris, France; Tehran, Iran; Milan and Salerno, Italy*

Background: Transthoracic echocardiography is the leading cardiac imaging modality for patients admitted with COVID-19, a condition of high short-term mortality. The aim of this study was to test the hypothesis that artificial intelligence (AI)–based analysis of echocardiographic images could predict mortality more accurately than conventional analysis by a human expert.

Methods: Patients admitted to 13 hospitals for acute COVID-19 who underwent transthoracic echocardiography were included. Left ventricular ejection fraction (LVEF) and left ventricular longitudinal strain (LVLS) were obtained manually by multiple expert readers and by automated AI software. The ability of the manual and AI analyses to predict all-cause mortality was compared.

Results: In total, 870 patients were enrolled. The mortality rate was 27.4% after a mean follow-up period of 230 ± 115 days. AI analysis had lower variability than manual analysis for both LVEF (P = .003) and LVLS (P = .005). AI-derived LVEF and LVLS were predictors of mortality in univariable and multivariable regression analysis (odds ratio, 0.974 [95% CI, 0.956-0.991; P = .003] for LVEF; odds ratio, 1.060 [95% CI, 1.019-1.105; P = .004] for LVLS), but LVEF and LVLS obtained by manual analysis were not. Direct comparison of the predictive value of AI versus manual measurements of LVEF and LVLS showed that AI was significantly better (P = .005 and P = .003, respectively). In addition, AI-derived LVEF and LVLS had more significant and stronger correlations to other objective biomarkers of acute disease than manual reads.

From the MedStar Health Research Institute and Georgetown University, Washington, District of Columbia (F.M.A.); Ultromics, Oxford, United Kingdom (T.D., R.U., G.M.W.); Experimental Therapeutics, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom (R.S.); University of Chicago, Chicago, Illinois (I.K., C.C.S., K.A., R.M.L.); Union Hospital, Tongji Medical College of HUST, Wuhan, China (M.X., C.W.); Philippine Heart Center, Quezon City, Philippines (E.S.T.); Radiology Institute of the University of São Paulo Medical School, São Paulo, Brazil (A.C.T.R.); Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran, Mexico City, Mexico (Z.Y.V.-O.); King's College Hospital, London, United Kingdom (M.J.M.); Centro Medico Nacional 20 de Noviembre, ISSSTE, Mexico City, Mexico (B.A.O.S.); Saint Antoine and Tenon Hospital, AP-HP, INSERM UMRS-ICAN 1166 and Sorbonne Université, Paris, France (L.S.-D.); Rajaie Cardiovascular Medical and Research Center, Echocardiography Research Center, Iran University of Medical Science, Tehran, Iran (A.A.); Baharloo Hospital, Tehran University of Medical Sciences, Tehran, Iran (A. Mostafavi); De Gasperis Cardio Center, Niguarda Hospital, Milan, Italy (A. Moreo); University of Salerno, Salerno, Italy (R.C.); and Northwestern University, Chicago, Illinois (A.N.).

This work was supported by the American Society of Echocardiography Foundation, the University of Chicago, and MedStar Health, with in-kind support from Ultromics.

T.D., R.U., and G.M.W. are full-time employees of Ultromics. R.S. has received consulting fees from Ultromics and Weatherden. M.J.M. is an advisory board and speaker's bureau member for Bracco and Philips. R.M.L. is an advisory board and speaker's bureau member for Philips; and is an advisory board for Caption Health. F.M.A. has received institutional (MedStar Health) research grants from TomTec, Ultromics, GE, and Caption Health; and is an unpaid scientific advisory board member for Ultromics.

A full list of additional WASE-COVID investigators is provided in the Appendix.

Reprint requests: Federico M. Asch, MD, FASE, MedStar Health Research Institute, 100 Irving Street, NW, Suite EB 5123, Washington, DC 20010 (E-mail: *federico.asch@medstar.net*).

0894-7317/$36.00

https://doi.org/10.1016/j.echo.2022.07.004

Conclusions: AI-based analysis of LVEF and LVLS had similar feasibility as manual analysis, minimized variability, and consequently increased the statistical power to predict mortality. AI-based, but not manual, analyses were a significant predictor of in-hospital and follow-up mortality. (J Am Soc Echocardiogr 2022;■:■-■.)

Keywords: Echocardiography, Machine learning, Artificial intelligence, Outcomes prediction, WASE, COVID-19, Left ventricular function

Abbreviations

4CH = Four-chamber

AI = Artificial intelligence

BNP = Brain natriuretic peptide

ICU = Intensive care unit

LV = Left ventricular

LVEF = Left ventricular ejection fraction

LVLS = Left ventricular longitudinal strain

ML = Machine learning

TTE = Transthoracic echocardiographic

WASE = World Alliance Societies of Echocardiography

Although it is considered mostly a respiratory disease, multiple organ systems are affected in patients with COVID-19. Growing evidence suggests that COVID-19-related cardiovascular complications play a significant role in disease severity and patient outcomes, including a higher risk for death.[1-3] Myocardial damage may occur because of a direct insult (myocardial infarction, thrombosis, myocarditis) or as a result of a systemic inflammatory response and may affect both the left and right ventricles.[4-6]

Transthoracic echocardiographic (TTE) imaging is the leading cardiac imaging modality for patients admitted with COVID-19,[7] as it can evaluate the full spectrum of cardiac involvement and can be performed safely in various settings, such as at the bedside in the emergency department or the intensive care unit (ICU).[6,8] Moreover, as myocardial injury has been linked with poor outcomes,[9] an echocardiogram obtained at admission may prove to be a powerful tool to predict outcomes in patients admitted with acute COVID-19. Today, the interpretation of echocardiographic images is based on manual analysis associated with considerable measurement variability, which is likely to affect its predictive value.

The use of artificial intelligence (AI) machine learning (ML)–based technologies in cardiovascular medicine is rapidly growing, resulting in increased automation of image processing and diagnostic interpretation. So far, most research has been focused on big data analysis through accessing large data sets, building models and algorithms to identify diagnostic patterns or to predict outcomes.[10,11] The role of AI in cardiovascular imaging and specifically echocardiography is expanding to facilitate image acquisition and analysis.[12-15] Moving toward a fully automated, AI-based analysis will result in lower variability of results than those obtained from reader-dependent techniques widely practiced today.[13] With lower variability and increased interpretation consistency, it is foreseeable that the use of automated measurements could improve the capacity to predict outcomes.[16] However, few studies have performed direct head-to-head comparisons between AI and conventional human interpretation of echocardiographic images.

The international World Alliance Societies of Echocardiography (WASE) COVID-19 Study was designed to describe echocardiographic characteristics and to identify parameters that would be prognostic of clinical outcomes in patients admitted with acute COVID-19. In this analysis, we aimed to test the performance of an AI ML-derived algorithm for the prediction of outcomes in patients admitted for acute COVID-19 and its incremental value to that of expert echocardiographer analysis. Specifically, we hypothesized that automated left ventricular (LV) function analysis obtained using the ML algorithms would have less interreader variability than expert readers, translating into better prediction of mortality.

METHODS

Study Design and Data Collection

The WASE COVID-19 Study enrolled adult patients admitted for acute COVID-19 (including positive antigen or polymerase chain reaction test results) during the first wave of the pandemic (January to September 2020). Patients were included if TTE imaging was performed during the initial COVID-19 hospitalization, and enrollment was performed in a prospective and retrospective manner. All follow-up was performed in a prospective manner. Patients were enrolled at 13 medical centers in nine countries worldwide. Because of differences in safety protocols to protect the acquiring operators, TTE examinations were ordered and acquired on the basis of local clinical practices and included both comprehensive and limited studies, acquired using various imaging equipment.[8,17,18] If patients underwent more than one TTE study, only the initial one was used. TTE studies included, at a minimum, a four-chamber (4CH) view, which was required for calculation of LV ejection fraction (LVEF), LV end-systolic and end-diastolic volumes, and LV longitudinal strain (LVLS), although two-chamber views were also used to determine biplane LVEF or average LVLS values, whenever available. LVLS was calculated as the average of all available segments from the 4CH and two-chamber views, as a long-axis view was not obtained in the vast majority of cases. The protocol was approved by each local ethics committee, and patients provided informed consent for any prospective clinical encounter or image acquisition. The in-hospital course and outcomes of the WASE COVID-19 Study are reported elsewhere.[19]

Digital Imaging and Communications in Medicine images were web-transferred to a cloud-based secure storage system that includes automated analysis of LVEF, LV volumes, and LVLS (EchoGo Core; Ultromics). Clinical information including demographic data, medical history, vital signs, and serum biomarkers was collected by local investigators and stored in a secure web-based system (Castor EDC; Castor). Biomarkers were collected whenever deemed clinically appropriate within 72 hours of echocardiographic acquisition and included brain natriuretic peptide (BNP) and C-reactive protein. To account for different biomarker assays used at each center, the level of each biomarker (BNP and C-reactive protein) was classified as either normal, borderline abnormal (<2 times the upper limit of normal), or abnormal (>2 times the upper limit of normal), on the basis of the reference values for each center. In addition to in-hospital

HIGHLIGHTS

- AI automated and manual quantification of LVEF and LVLS had similar feasibility.
- AI reduced variability, increasing power to predict outcomes compared with manual.
- AI-based, but not manual, LV analyses predicted mortality in patients with COVID-19.

clinical outcomes, outpatient follow-up was performed ≥3 months after hospital admission by review of medical records, office encounter, and/or phone call. The primary outcome of the WASE COVID-19 Study was all-cause mortality (in-hospital and up to 6 months of follow-up).

Image Analysis

Each TTE study underwent two independent forms of LV analysis, with multiple independent runs of each: (1) cloud-based and automated and (2) conventional human reads by board-certified experts. Each of these reads was performed blinded to each other.

Automated LV analyses were performed using an AI ML-based algorithm (EchoGo Core), which contoured the LV endocardium automatically to enable Simpson's calculation of LV end-diastolic volume and end-systolic volume and LVEF, as well as speckle-tracking-based LVLS. EchoGo Core is a cloud-based, vendor-neutral program that uses AI to automatically contour the LV endocardium and cannot be manually adjusted. The software is commercially available and has been recently cleared by the US Food and Drug Administration. Further details on the AI architecture and development are provided in the Supplemental Appendix. Briefly, the software automatically classifies the echocardiographic views, which are confirmed by an operator for quality control. The best cardiac cycles and frames are automatically identified to finally calculate LV end-diastolic volume and end-systolic volume, LVEF, and LVLS. Operators could only accept or reject the final clip (loop) to report but could not edit the LV tracings or cycle and frame selections. If a frame was rejected, the software would select a new clip or frame, which again would have to be accepted or rejected by the operator. Each TTE study was analyzed on two separate occasions by different operators to test reproducibility in rejecting or accepting automated tracing and measurements. The two operators for each TTE study were randomly assigned from a pool of 11 operators. Processed studies were accepted or rejected on the basis of processing quality (view selection, contouring success, Digital Imaging and Communications in Medicine conformance, etc).

All measurements were repeated in three rounds of manual quantifications by independent experts blinded to any clinical status, following conventional methodology. The three readers for each TTE study were randomly assigned from a pool of eight experts, also present in the AI operator pool, all board-certified echocardiographers. For each case, manual rounds 1 and 2 were analyzed ≥30 days apart by the same operator to determine the intraobserver variability, while round 3 was analyzed by a different operator to determine interobserver variability. All LV volumes and LVEFs were obtained by performing endocardial tracings and using the method of disks (modified Simpson's rule).[20] Only cases with acceptable quality LV views were included, which was defined as a lack of apical foreshortening and adequate visualization of all segments in the apical 4CH view. Overall, 67.2% of the AI-based parameters (LV volumes, LVEF, and LVLS) and 70.4% of the manual parameters were obtained using the biplane method, while the rest were obtained from the 4CH view alone. For each read, the end-diastolic and end-systolic frames (largest and smallest LV volumes) across image clips selected by the operators were recorded to determine the variability associated with image contouring alone in contrast to difference in frames and clips, where acquisition variability may additionally contribute.

Statistical Analysis

All statistical analysis was performed with R version 4.0.4. Continuous variables are expressed as mean ± SD or as median (interquartile range), according to the data distribution, and compared using Student's *t* test or the Wilcoxon rank sum tests, as appropriate. Categorical data, presented as numbers and percentages, were compared using the χ^2 test. Biochemical markers underwent natural logarithmic transformation. Cox proportional hazard regression and binomial generalized linear models with logit function were performed on the mean manual values and AI values to evaluate the univariable relationship between echocardiographic parameters and in-hospital and 30-day mortality. Date of death during outpatient follow-up was not available in some cases, which affected the hazard proportionality beyond 30 days, and therefore only linear regression was used for the follow-up analysis. Results from regression models are reported as odds ratios with 95% CIs, which were analyzed as continuous variables in 1% increments both for LVEF and LVLS. Univariate survival analysis was performed for time to death, using Cox proportional hazard regression. Forest plots and cumulative hazard plots were constructed for visualization. For direct head-to-head comparison of AI and manual measurements for prediction of death, we compared the increase in prognostic value directly through likelihood ratio test comparing both prognostic models for goodness of fit.

To assess the variability factors associated with the quantification of echocardiograms, a general linear mixed model was used to determine the within-patient variability components attributed to operator, frame selection, and image quality, included as random effects. Inter- and intraoperator variability was assessed using Pearson correlations and intraclass correlation coefficients for all cases, as well as subset for instances in which operators processed the same or different end-diastolic and end-systolic frames (i.e., cycles/clips). Operator influence on the variability in LVEF and LVLS for manual and AI measurements was visualized using principal-component analysis. Further details on the use of principal-component analysis are described in the Supplemental Appendix. Levene tests were used to assess whether the observed difference in variability within each variable was significantly different between manual and AI contouring. The effect size, power, and sample size calculations were used to assess the power the observed variability had on predicting clinical outcomes. A correlation matrix was performed to compare multivariable correlations of serum biomarkers, blood pressure, LVEF, and LVLS; this was performed by constructing a network using the Pearson correlation coefficient and Bonferroni correction for multiple testing. For all statistical analysis, significance was set at $P < .05$.

RESULTS

Over a 9-month period (January to September 2020), the WASE COVID-19 Study enrolled 870 patients from 13 centers in nine

Table 1 Demographic characteristics of all patients in the study, those in whom there was a manual read or an AI read, and those in whom both manual and AI reads were available

	All patients (*N* = 870)	Manual reads (*n* = 699)	AI reads (*n* = 511)	Both AI and manual reads (*n* = 476)
Patient demographics				
Age, y	59.38 ± 15.07	59.58 ± 15.00	59.94 ± 15.03	60.06 ± 14.94
Sex				
Female	381 (43.8)	303 (43.3)	229 (44.8)	210 (44.1)
Male	488 (56.1)	395 (56.5)	281 (55.0)	265 (55.7)
Unknown	1 (0.1)	1 (0.1)	1 (0.2)	1 (0.2)
Ethnicity				
White non-Hispanic	197 (22.6)	153 (21.9)	125 (24.5)	121 (25.4)
White Hispanic	152 (17.5)	110 (15.7)	83 (16.2)	72 (15.1)
Black	136 (15.6)	111 (15.9)	98 (19.2)	92 (19.3)
Asian	271 (31.1)	230 (32.9)	137 (26.8)	130 (27.3)
Mixed	72 (8.3)	64 (9.2)	53 (10.4)	47 (9.9)
Other	34 (3.9)	25 (3.6)	13 (2.5)	12 (2.5)
Unknown	8 (0.9)	6 (0.9)	2 (0.4)	2 (0.4)
Clinical parameters				
Blood pressure				
SBP, mm Hg	123.3 ± 19.30	124.2 ± 19.12	126.5 ± 19.13	127 ± 19.18
DBP, mm Hg	74.57 ± 12.15	74.86 ± 12.30	75.45 ± 12.09	75.68 ± 12.20
Heart rate, beats/min	85.26 ± 15.46	84.32 ± 15.25	84.95 ± 15.22	84.71 ± 14.99
Status at initial TTE study				
ICU	402 (46.2)	316 (45.2)	216 (42.3)	201 (42.2)
Mechanical ventilation	236 (27.1)	182 (26.0)	116 (22.7)	107 (22.5)
Hemodynamic support	155 (17.8)	120 (17.2)	74 (14.4)	69 (14.5)
Previous conditions				
Heart disease	544 (62.5)	438 (62.7)	304 (59.5)	286 (60.1)
Lung disease	127 (14.6)	98 (14.0)	72 (14.1)	65 (13.7)
Kidney disease	80 (9.2)	65 (9.3)	49 (9.6)	48 (10.1)
Hypoxemia	24 (2.8)	17 (2.4)	11 (2.2)	11 (2.3)
Biomarkers				
BNP				
Abnormal	160 (18.4)	131 (18.7)	97 (19.0)	94 (19.7)
Borderline	46 (5.3)	40 (5.7)	32 (6.3)	31 (6.5)
Normal	153 (17.6)	121 (17.3)	104 (20.4)	98 (20.6)
Not measured	511 (58.7)	407 (58.2)	278 (54.4)	253 (53.2)
CRP				
Abnormal	635 (73.0)	501 (71.7)	371 (72.6)	344 (72.3)
Borderline	51 (5.9)	37 (5.3)	26 (5.1)	23 (4.8)
Normal	106 (12.2)	92 (13.2)	70 (13.7)	66 (13.9)
Not measured	78 (9.0)	69 (9.9)	44 (8.6)	43 (9.0)
Outcome				
Death (in-hospital)	188 (21.6)	152 (21.7)	98 (19.18)	91 (19.36)
Death (follow-up)	238 (27.4)	192 (27.5)	132 (25.8)	123 (26.2)

CRP, C-reactive protein; *DBP*, diastolic blood pressure; *SBP*, systolic blood pressure.
Data are expressed as mean ± SD or as number (percentage).

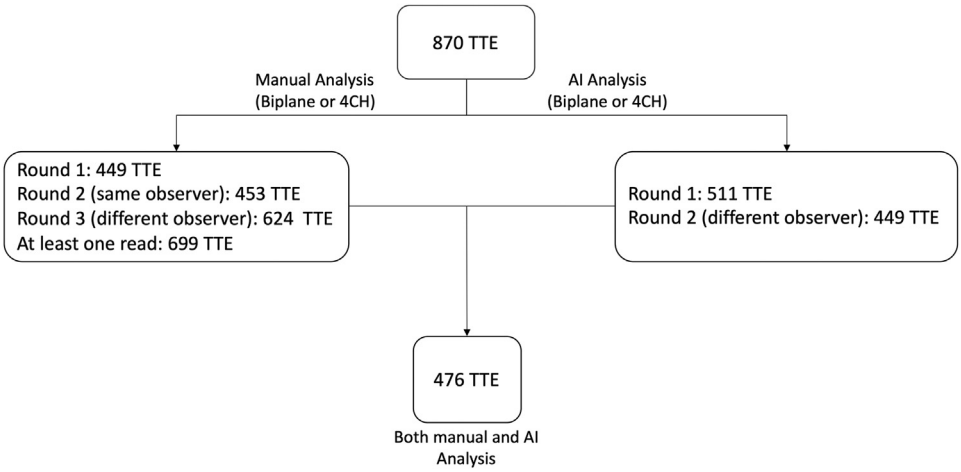


Figure 1 Flowchart describing feasibility of analysis in each round. Manual reads were performed by randomly selected operator from a pool of seven experts. Manual rounds 1 and 2 were performed blindly by the same operator to derive intraobserver variability. Round 3 was performed by a different operator to derive interobserver variability. AI analysis was performed in two separate rounds to test consistency in selection of the specific cardiac cycle and to test intraobserver variability. A total of 476 echocardiograms were successfully analyzed both in at least one manual and one AI run.

countries (Table 1). By protocol design, all patients were hospitalized at the time of the initial TTE examination, 46.2% were admitted to an ICU, 27.1% were receiving mechanical ventilation, and 17.9% were on hemodynamic support (inotropic drugs, vasopressors, intra-aortic balloon pump, or LV assist device). TTE studies were obtained a median of 3 days after admission (interquartile range, 1-9 days). The mean time to follow-up was 230 ± 115 days. Overall, 238 patients (27.4%) died at time of final follow-up (≥3 months from the time of COVID-19 admission), 188 (21.6%) during the initial hospitalization, and 50 (5.7%) during subsequent outpatient follow-up (Table 1).

Out of the 870 echocardiograms obtained, 449 (52%), 453 (52%), and 624 (72%) cases were successfully manually contoured in manual rounds 1, 2, and 3, respectively. Overall, 699 cases (80%) were manually contoured in at least one of the three manual rounds (Figure 1). AI-based contouring was performed on all 870 cases, and the final contours were approved or rejected by a trained operator. In the first AI round, 511 cases (59%) were approved for analysis (reasons for missing analysis were as follows: 166 were considered by the operator to be of poor image quality, 43 were missing the needed views, and the rest had image formatting incompatibilities or other technical problems). In the second AI round (performed by different operators), 449 cases were approved. A total of 476 studies (54.7%) were successfully analyzed by both human experts and AI (Figure 1). For the manual quantification, different cycles and end-diastolic and end-systolic frames were selected by the operators in 305 cases compared with 336 with different operators using the AI program (Table 2).

Differences between Manual and AI Contouring

There was substantial overlap in the frequency distribution histograms of LVEF and LVLS measurements between manual and AI analysis. The intermethod analysis demonstrated a mean difference of −1.756 (95% CI, −2.704 to −0.808; $P < .001$) for LVEF and

Table 2 Interoperator agreement using manual or AI-based analysis and dependent on frame selection

Method	Measure	Frame selection	*n*	*R* (Pearson correlation) (95% CI)	ICC (95% CI)	Coefficient of variation, %
AI	LVEF	All	385	0.853 (0.824-0.878)	0.854 (0.824-0.879)	10.74
Manual			319	0.670 (0.605-0.727)	0.655 (0.573-0.722)	19.74
AI	LVEF	Same	49	0.996 (0.994-0.998)	0.996 (0.993-0.998)	
Manual			14	0.683 (0.239-0.891)	0.680 (0.240-0.886)	
AI	LVEF	Different	336	0.832 (0.796-0.862)	0.832 (0.796-0.862)	
Manual			305	0.671 (0.504-0.728)	0.654 (0.569-0.723)	
AI	LVLS	All	385	0.789 (0.784-0.824)	0.789 (0.748-0.824)	19.15
Manual			339	0.430 (0.336-0.515)	0.430 (0.336-0.515)	39.95
AI	LVLS	Same	49	0.987 (0.977-0.993)	0.987 (0.977-0.993)	
Manual			14	0.497 (<0.001-0.813)	0.510 (<0.001-0.814)	
AI	LVLS	Different	296	0.761 (0.712-0.803)	0.761 (0.712-0.803)	
Manual			305	0.427 (0.330-0.514)	0.426 (0.330-0.514)	

ICC, Intraclass correlation coefficient.

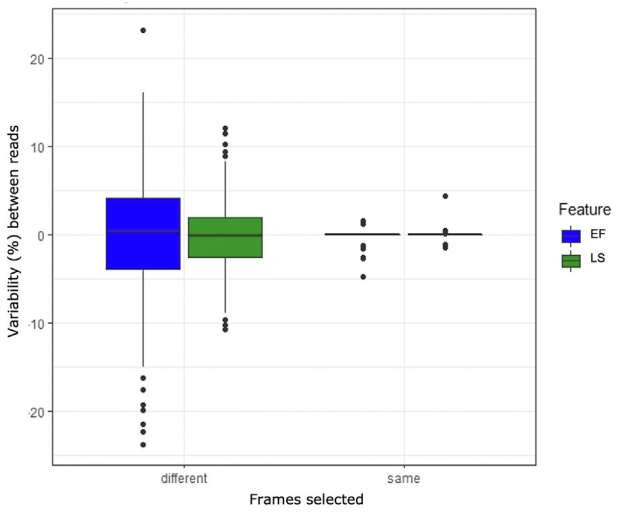


Figure 2 AI interreader variability according to frame selection. The *vertical axis* demonstrates variation from read 1 to read 2. Interreader variability in LVEF and LVLS was larger when there was discordance in frame selection for the measurements *(left plots)*. When the same frame was selected for measurement of LVEF and LVLS, variability was minimal *(right plots)*.

−1.614 (95% CI, 2.140 to −1.087; $P < .001$) for LVLS. Interobserver intraclass correlation coefficients for manual and AI reads are provided in Table 2, demonstrating lower interoperator reproducibility for manual compared with AI reads. This reproducibility was further reduced when operators chose different end-diastolic and end-systolic frames compared with when they selected the same frames. When choosing the same frames, the AI reads showed near perfect interoperator reproducibility, which was reduced when choosing different cycles, although it remained substantially better than for the manual reads (Figure 2). Using a general linear mixed model, the variability contributions attributed to operator, image quality, and cycle frame selection were calculated for manual and AI reads of LVEF and LVLS (Table 3). Frame selection contributed minimally to interoperator variability both for the manual and AI reads. On sensitivity analysis, a one-frame difference in diastolic contouring represented a mean LVEF change of 0.5 ± 5.2% and a mean LVLS change of −0.4 ± 2.8% (a shift up to two frames did not change the associations with outcomes). However, the variability attributed to the operator in the manual reads (47.4% and 51.8% of the total variability for LVEF and LVLS, respectively) was disproportionately higher compared with the variability attributed to the operator in the AI reads (0.18% and 1.42% of the total variability). Image quality did not contribute to the observed variability.

Given that the operators represented the greatest source of variability by general linear mixed model analysis, operator influence on the variability in LVEF and LVLS in manual and AI measurements was visualized using principal-component analysis (Supplemental Figure 1). This showed improved clustering of datapoints by AI analysis compared with the manual reads. The difference in variability between manual and AI analysis was significant using the Levene test for both LVEF ($P = .003$, $F = 8.898$, $df = 1$) and LVLS ($P = .005$, $F = 7.982$, $df = 1$), which attributed to a higher statistical power. Using the variability components for manual and AI reads, manual reads would require a 1.8-fold increase in sample size to achieve the same power ($1 - \beta = 0.8$) as AI reads. Correlations between LVEF or LVLS and other biomarkers were stronger for AI-based than manual contouring, but they were low overall (Supplemental Table 2).

Predicting Mortality with AI versus Manual LV Analysis

Univariable logistic regression showed that both LVEF and LVLS were significant predictors of mortality when measured with AI contouring but not when done manually by experts (Supplemental Table 1). This was true for both in-hospital and overall mortality (Table 4) and when only the 4CH views were used (single-plane analysis). LV volumes, on the other hand, failed to show predictive value when obtained by either manual or AI analysis. AI-derived LVEF and LVLS showed significant association with mortality both in-hospital and at final follow-up and when analyzed as a continuous or categorical variable. For the manual reads, only LS analyzed as a categorical variable (cutoff = −16%, obtained from receiver operating characteristic analysis) was a predictor of in-hospital mortality. In univariable logistic regression, multiple additional variables showed significant predictive value for mortality both in-hospital and through the final follow-up: age, admission to the ICU, requiring mechanical ventilation or hemodynamic support, previous heart disease or lung disease, C-reactive protein, and BNP (Table 4).

When including the human or AI echocardiographic measurements in a forward-step multivariable logistic regression that independently selected BNP as a covariate when restricted to LVEF inclusion, and BNP and need for mechanical ventilation as covariates when restricted to LVLS inclusion, a proportionally higher increase in odds ratio (albeit fairly modest) was apparent for AI-based

Table 3 Within-patient variability across manual and AI reads

	LVEF		LVLS	
	Manual	AI	Manual	AI
Variable	Variability (% total)	Variability (% total)	Variability (% total)	Variability (% total)
Frame	1.033 (1.40)	2.362 (6.30)	0.876 (2.74)	0.588 (5.96)
Operator	34.946 (47.39)	0.067 (0.18)	16.537 (51.81)	0.140 (1.42)
Reading round	<0.0001 (<0.001)	0.016 (0.04)	0.115 (0.36)	0.109 (1.11)
Image quality	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)

Using a general linear mixed model, variability components for random nested effect were calculated and described. Variability is expressed as a percentage of the total.

Table 4 Univariable logistical regression against outcomes across AI and manual reads

Parameter	Mortality			
	In-hospital		Follow-up	
	Odd ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*
Echocardiographic parameters (continuous)				
LVEF manual	0.985 (0.969-1.003)	.083	0.990 (0.975-1.005)	.187
LVEF AI	0.970 (0.952-0.988)	.001	0.974 (0.956-0.991)	.003
LVLS manual	1.035 (0.999-1.074)	.058	1.024 (0.991-1.059)	.155
LVLS AI	1.082 (1.035-1.132)	<.001	1.060 (1.019-1.105)	.004
LVESV manual*	1.085 (0.806-1.456)	.588	1.050 (0.799-1.378)	.724
LVESV AI*	1.289 (0.935-1.771)	.118	1.097 (0.801-1.495)	.558
LVEDV manual*	1.087 (0.810-1.454)	.575	1.050 (0.799-1.378)	.724
LVEDV AI*	1.073 (0.675-1.700)	.876	1.966 (0.622-1.493)	.877
Echocardiographic parameters (categorical)				
LVEF manual (reference <60%)	0.729 (0.457-1.159)	.182	0.729 (0.457-1.159)	.182
LVEF AI (reference <60%)	0.452 (0.282-0.722)	.001	0.479 (0.311-0.736)	.001
LVLS manual (reference <−16%)	2.061 (1.268-3.334)	.003	2.061 (1.268-3.334)	.003
LVLS AI (reference <−16%)	2.616 (1.833-4.208)	<.001	1.887 (1.223-2.911)	.004
Significant clinical parameters				
Age	1.030 (1.013-1.048)	<.001	1.026 (1.012-1.042)	<.001
Status at initial TTE study				
ICU	6.139 (3.650-10.708)	<.001	3.777 (2.441-5.915)	<.001
Ventilator	10.800 (6.421-18.491)	<.001	7.215 (4.422-11.951)	<.001
LV support	7.080 (4.054-12.504)	<.001	6.295 (3.583-11.334)	<.001
Previous conditions				
Heart disease	1.907 (1.160-3.216)	.013	1.540 (0.989-2.429)	.059
Lung disease	1.952 (1.065-3.488)	.0263	1.391 (0.722-2.290)	.370
Biomarkers				
CRP (reference normal)				
Borderline	1.157 (0.055-9.714)	.902	8.625 (1.785-63.116)	.013
Abnormal	6.956 (2.484-29.042)	.001	11.611 (3.497-71.959)	<.001
BNP (reference normal)				
Borderline	2.115 (0.596-6.911)	.221	1.962 (0.726-5.138)	.173
Abnormal	4.433 (1.971-11.017)	<.001	2.333 (1.188-4.715)	.016
DBP	0.959 (0.934-0.983)	<.001	0.974 (0.952-0.995)	.016

CRP, C-reactive protein; *DBP*, diastolic blood pressure; *LVEDV*, LV end-diastolic volume; *LVESV*, LV end-systolic volume.
Only parameters with *P* values < .05 in univariate logistic regression (binomial with logit link) are included. Odds ratios were analyzed as continuous variable in 1% increments.
*Log_2-transformed values.

measurements compared with those produced by manual human analysis (Table 5). As expected, AI LVLS values (but not manual values) were a significant predictor in the nonventilated group. In the ventilated group, however, prediction was dominated by the ventilation factor for the LVLS models, rendering in no significant addition in prediction of the other parameters.

Results of multivariable logistic regression models using AI and manual measurements correcting for other clinical variables are presented in Table 6.

The Cox regression analysis for in-hospital mortality (Supplemental Table 3) produced results similar to those observed with the logistic regression, with AI reads showing significant hazards compared with manual reads (Figure 3). Further cumulative hazards for cases read manually or with AI for in-hospital mortality are shown in Figure 4.

For direct head-to-head comparison of AI versus manual measurements, the increase in prognostic value was compared directly using a likelihood ratio test comparing both prognostic models for goodness of fit to all-cause mortality. In this analysis, the model with the higher log likelihood is the one that fits the outcome better. The log likelihood was −220.92 for manual LVEF and −217.19 for AI contouring ($P = .005$), indicating better goodness of fit for the AI model. Similarly, the log likelihood for the LVLS manual model was −220.54 and −216.21 for AI contouring ($P = .003$), also indicating better goodness of fit for the AI model.

Table 5 Multivariable forward-step logistical regression for outcomes by AI and manual reads

	Model 1 (LVEF manual)		Model 2 (LVEF AI)		Model 3 (LVLS manual)		Model 4 (LVLS AI)	
Parameter	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
LVEF manual	0.992 (0.967-1.018)	.532						
LVEF AI			0.971 (0.945-0.997)	.028				
LVLS manual					1.038 (0.975-1.108)	.254		
LVLS AI							1.096 (1.022-1.179)	.012
BNP								
Borderline	2.069 (0.581-6.776)	.236	1.795 (0.498-5.951)	.346	1.238 (0.317-4.395)	.746	0.909 (0.214-3.448)	.892
Abnormal	3.998 (1.664-10.472)	.003	3.134 (1.292-8.209)	.014	2.896 (1.120-8.026)	.033	2.662 (1.073-7.093)	.040
Mechanical ventilation					6.927 (3.000-16.500)	<.001	7.582 (3.202-18.712)	<.001
In patients on mechanical ventilation								
LVLS manual					0.980 (0.866-1.105)	.714		
LVLS AI							1.093 (0.967-1.260)	.178
BNP								
Borderline					2.391 (0.317-23.200)	.410	1.571 (0.185-16.350)	.683
Abnormal					5.091 (0.785-47.10)	.108	3.951 (0.668-32.951)	.151
In patients not on mechanical ventilation								
LVLS manual					1.064 (0.988-1.154)	.116		
LVLS AI							1.096 (1.006-1.201)	.042
BNP								
Borderline					0.576 (0.029-3.685)	.621	0.574 (0.029-3.690)	.619
Abnormal					2.323 (0.777-7.542)	.140	2.359 (0.816-7.408)	.121

The step-wise method uses Akaike information criterion metrics to build the stepwise model. Odds ratios were analyzed as continuous variable in 1% increments.

DISCUSSION

In this study, we applied AI-based technology to perform automated echocardiographic analysis of LV function. We have shown that quantifying LV systolic function with AI was feasible in a similar proportion of cases to manual contouring and that AI contours had less variability, although in many cases in our study it was performed solely from a 4CH view (32.8% and 29.6% for AI and manual measurements, respectively) because of limited acquisition (as opposed to the recommended biplane or three-plane methods). Furthermore, the use of AI increased the statistical power of both LVEF and LVLS to predict all-cause mortality in hospitalized patients with COVID-19 across different health care settings, using different TTE platforms, and in a wide spectrum of image quality. Our findings highlight the role of AI-assisted echocardiography in prediction of outcomes and how it complements the prognostic role of other important clinical variables such as requirement of mechanical ventilation.

Variability in Echocardiographic Analysis

The current American Society of Echocardiography guidelines recommend that LVEF be measured using the Simpson biplane method of disks, which involves manual tracing of the LV borders in both apical 4CH and two-chamber views.[20] However, quantification by echocardiography is susceptible to significant inter- and intraobserver variability, because of inherent subjectivity in endocardial border delineation.[21] Variability in measurements between readers can occur because of beat-to-beat variations in LV size and shape (i.e., respiratory cycle or arrhythmia), as well as reader bias resulting from knowledge of patient diagnosis, condition, or previous results. To address variability among readers, there has been growing interest in the development of automated software tools.[12,22,23] In this study, the major source of variability was found to be differences in manual contouring between the operators. Although differences in acquisition technique (i.e., sonographer experience, echocardiographic equipment) could also account for variability, all readers in our study (including the AI software) were presented with the same echocardiograms, thus eliminating this potential source of inconsistency, while focusing on the analysis variability (specific loop selection and contouring). It is evident from our results that automated, AI-based analysis reduces variability almost exclusively to the selection of frames to be measured (within the same or different cardiac cycles), which is the only phase of this analysis in which humans can provide input while using this specific AI software (as the AI-based analysis has a human component, it is therefore not strictly automated). It is foreseeable

Table 6 Multivariate logistic regression models with in-hospital or follow-up death as outcome variables

Parameters	In-hospital death		Follow-up	
	OR (95% CI)	*P*	OR (95% CI)	*P*
Model A				
LVEF, manual	0.984 (0.964-1.004)	.121	0.988 (0.970-1.005)	.166
Age	1.041 (1.020-1.063)	<.001	1.031 (1.014-1.049)	<.001
Sex (reference: female)	0.945 (0.541-1.650)	.841	0.912 (0.561-1.483)	.709
ICU (reference: no)	2.085 (1.053-4.124)	.034	1.450 (0.800-2.585)	.213
Ventilation	7.623 (3.685-16.294)	<.001	5.118 (2.590-10.367)	<.001
Hemodynamic support	1.423 (0.674-2.969)	.350	1.824 (0.882-3.761)	.103
Model B				
LVEF, AI	0.974 (0.952-0.996)	.022	0.976 (0.957-0.996)	.017
Age	1.041 (1.020-1.064)	<.001	1.031 (1.014-1.050)	<.001
Sex (reference: female)	0.913 (0.521-1.601)	.749	0.893 (0.548-1.457)	.651
ICU (reference: no)	1.994 (1.004-3.952)	.047	1.397 (0.769-2.498)	.264
Ventilation	7.948 (3.827-17.027)	<.001	5.345 (2.696-10.856)	<.001
Hemodynamic support	1.308 (0.615-2.742)	.481	1.692 (0.814-3.506)	.156
Model C				
LVLS, manual	1.017 (0.975-1.062)	.450	1.014 (0.976-1.054)	.471
Age	1.041 (1.020-1.064)	<.001	1.031 (1.014-1.050)	<.001
Sex (reference: female)	0.942 (0.539-1.646)	.832	0.912 (0.561-1.484)	.711
ICU (reference: no)	2.127 (1.071-4.223)	.030	1.459 (0.802-2.616)	.209
Ventilation	7.452 (3.601-15.915)	<.001	5.089 (2.568-10.331)	<.001
Hemodynamic support	1.369 (0.645-2.861)	.408	1.761 (0.851-3.627)	.125
Model D				
LVLS, AI	1.056 (1.003-1.114)	.039	1.043 (0.997-1.092)	.072
Age	1.040 (1.019-1.063)	<.001	1.031 (1.014-1.049)	<.001
Sex (reference: female)	0.900 (0.510-1.572)	.700	0.892 (0.548-1.453)	.644
ICU (reference: no)	2.052 (1.037-4.057)	.038	1.428 (0.789-2.546)	.232
Ventilation	7.660 (3.691-16.430)	<.001	5.229 (2.640-10.623)	<.001
Hemodynamic support	1.308 (0.611-2.755)	.483	1.680 (0.806-3.481)	.163

Each model contains the variables age, sex, ICU admission, ventilation, and hemodynamic support together with either manual or AI-derived LVEF or LVLS. Odds ratios were analyzed as continuous variable in 1% increments.

that a combination of an AI analysis and AI-guided acquisition could further reduce variability.

In a multicenter study by Knackstedt *et al.*,[24] a fully automated analysis of two-dimensional echocardiograms provided both rapid and reproducible assessment of LVEF and LVLS. The investigators found significant differences in interobserver variability among study sites but no significant variability with respect to the automated LVEF tracings, findings similar to our study and to those in other fields, such as cardiac magnetic resonance or computed tomography.[25-27]

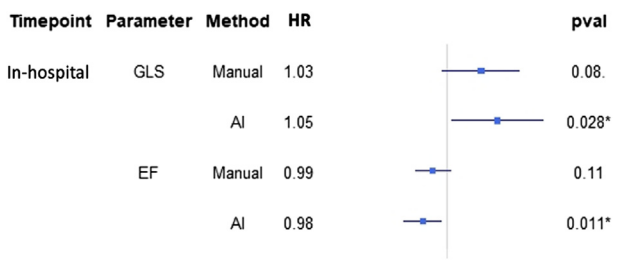


Figure 3 Forest plot for Cox proportional hazard regression against outcomes across AI and manual reads. *EF*, LV ejection fraction; *LS*, LV longitudinal strain.

Novel Uses of AI in Echocardiography

A recent study by Asch *et al.*,[12] using an unconventional algorithm for automated LVEF calculation, helped pave the way for our present study using a different, novel deep learning algorithm to analyze a large cohort of patients with acute COVID-19. The good performance of the AI analysis in our study was possible despite challenges with image quality, which is a common problem in patients with acute COVID-19 or other diseases in the ICU or on mechanical ventilation. In the European Association of Cardiovascular Imaging/American Society of Echocardiography Inter-Vendor Comparison Study, it was possible to analyze LVLS in 72% to 100% of studies, a higher proportion than in our study.[28] In this study, participants were scanned in a dedicated research setting, while in ours, all were inpatients with acute COVID-19, many intubated in an ICU setting with significant

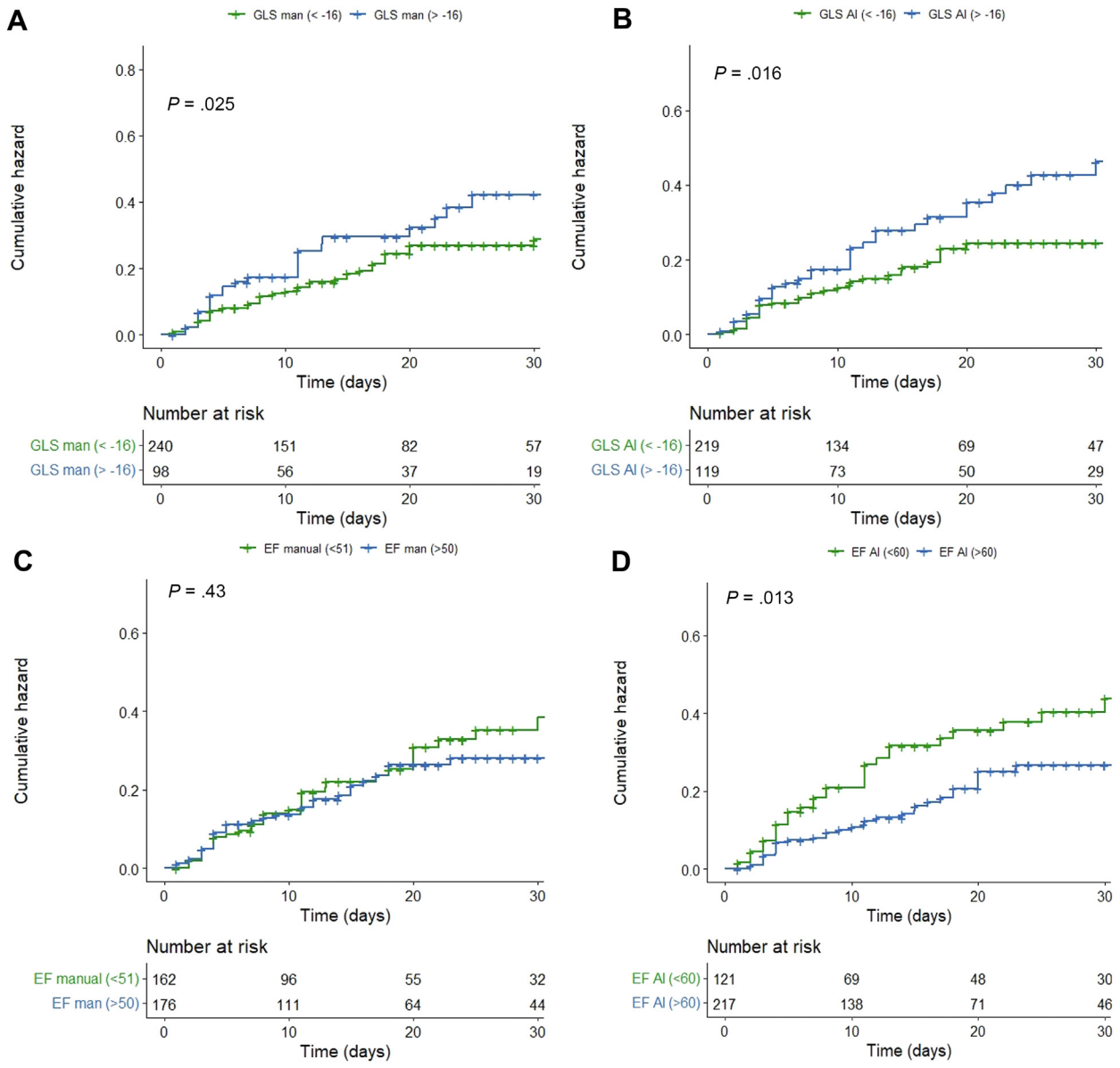


Figure 4 Kaplan-Meier cumulative hazards plots for Cox proportional hazard regression against in-hospital (<30-day) mortality across LVLS manual **(A)**, LVLS AI **(B)**, LVEF manual **(C)**, and LVEF AI **(D)** reads.

difficulties imposed by strict safety precautions.[18] As a result, most centers performed limited TTE studies, such that the median number of video loops was 32 (interquartile range, 19-42), many without electrocardiography.

The value of AI in cardiac imaging, however, goes beyond image analysis.[29,30] Recent advances have made automated image acquisition, segmentation, and interpretation possible across multiple imaging modalities, including computed tomography, cardiac magnetic resonance imaging, and echocardiography.[22,31] Automated LVEF and LVLS analysis and even disease identification are possible through the use of convolutional neural networks.[15,32,33] Lang *et al.*[13] recently described in healthy adults the high performance of a novel deep learning algorithm that automatically identifies and organizes images into "thematic stacks," while making automated measurements to accelerate and streamline the image review process.[13] In our study, AI software was able to reliably quantify LVEF and LVLS with less variability than manual expert readers.

Reduced Variability Improves Prediction of Outcomes

By reducing variability, prediction models can be more powerful and accurate than conventional interpretation by expert readers. AI contouring resulted in a marked increase in association to clinical outcomes, predicting mortality to a greater degree of accuracy and with increased correlations to other objective serum biomarkers. The predictive ability of AI-based measurements was additive and independent to that of other clinical variables in patients without mechanical ventilation. In the ventilated group, however, prediction was dominated by the ventilation factor. It is postulated that this is due to a more consistent or predictable behavior associated with AI

quantification, possibly subject to fewer human biasing factors. Until recently, AI-based algorithms for predicting outcomes from cardiac imaging were based on databases comprising static images or reports rather than video analysis.[10,11,34] More recently, however, the use of AI-assisted analysis of echocardiographic videos in a broad population has shown to be superior in predicting mortality than other clinical prediction models or the cardiologist's impression from echocardiographic and clinical data.[16] In a different study of nonacute asymptomatic patients with risk factors for heart failure, automated LVLS measurement was not superior to semiautomated analysis in predicting future cardiac events.[35] Our study is unique in that the improved predictive ability was achieved with a simple AI output of a single echocardiographic variable (LVEF or LVLS) and was further improved by combination with biomarkers. To our knowledge, this is the first report on the use of AI-based automated echocardiographic analysis for prediction of mortality in patients with acute COVID-19. Our study highlights that more sensitive results can be obtained using AI compared with manual measurements, which does not conflict with prior data supporting the value of LVLS as a predictive tool in other disease states.

Limitations

The main limitations of the study are that patients were enrolled in a retrospective manner, with no echocardiographic standardized acquisition. In addition, although image analysis was standardized, not all echocardiograms could be quantified. Although these findings may be applicable to patients with COVID-19, they do not necessarily apply to other disease states or other AI technologies. However, if these findings were broadened to a wider patient population with better image quality, it is conceivable that AI contouring could be feasible in a much higher proportion of patients. Were this to be the case, then AI could increase statistical power to predict outcomes, possibly requiring smaller sample sizes in clinical trials. The application of this methodology should thus be further assessed in more typical cardiology patient cohorts and clinical settings. Finally, we acknowledge that our study was a focused evaluation of LV variables and that we were looking at the relative contribution of these variables versus the clinical course of the patient, which is an important consideration in the setting of this highly ill population.

CONCLUSION

Automated quantification of LVEF and LVLS using AI in the WASE COVID-19 Study had similar feasibility to manual contouring, minimized variability, and consequently increased the statistical power to predict mortality. AI-based but not manual analyses were significant predictors of in-hospital and follow-up mortality. Application of this technology to other diseases or clinical trials may increase the accuracy of predicting outcomes or detecting clinical changes over time.

ACKNOWLEDGMENTS

We thank Victor Mor-Avi of the University of Chicago; Katie Ions, Nancy Spagou, Alex Hudson, Jake Kenworthy, Lokken Wong, Angela Mumith, Will Hawkes, and Louise A. Tetlow of Ultromics; Andrea Van Hoever of the American Society of Echocardiography; and Patricia Marques, Gaynor Jones, Seamus Walker, Gurpreet Dhadday, Aiko Nepomuceno, Jherick Tay, Claire Mitchell, Sarah Hastings, Emma-Jayne Robins.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://doi.org/10.1016/j.echo.2022.07.004.

REFERENCES

1. Lala A, Johnson KW, Januzzi JL, Russak AJ, Paranjpe I, Richter F, et al. Prevalence and impact of myocardial injury in patients hospitalized with COVID-19 infection. J Am Coll Cardiol 2020;76:533-46.
2. Uriel N, Sayer G, Clerkin KJ. Myocardial injury in COVID-19 patients: the beginning or the end. J Am Coll Cardiol 2020;76:547-9.
3. Shi S, Qin M, Shen B, Cai Y, Liu T, Yang F, et al. Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. JAMA Cardiol 2020;5:802-10.
4. Abbasi J. Researchers investigate what COVID-19 does to the heart. JAMA 2021;325:808-11.
5. Li Y, Li H, Zhu S, Xie Y, Wang B, He L, et al. Prognostic value of right ventricular longitudinal strain in patients with COVID-19. JACC Cardiovasc Imaging 2020;13:2287-99.
6. Szekely Y, Lichter Y, Taieb P, Banai A, Hochstadt A, Merdler I, et al. Spectrum of cardiac manifestations in COVID-19: a systematic echocardiographic study. Circulation 2020;142:342-53.
7. Citro R, Pontone G, Bellino M, Silverio A, Iuliano G, Baggiano A, et al. Role of multimodality imaging in evaluation of cardiovascular involvement in COVID-19. Trends Cardiovasc Med 2021;31:8-16.
8. Kirkpatrick JN, Grimm R, Johri AM, Kimura BJ, Kort S, Labovitz AJ, et al. Recommendations for echocardiography laboratories participating in cardiac point of care cardiac ultrasound (POCUS) and Critical care echocardiography training: report from the American Society of Echocardiography. J Am Soc Echocardiogr 2020;33:409-22.e4.
9. Shi S, Qin M, Cai Y, Liu T, Shen B, Yang F, et al. Characteristics and clinical significance of myocardial injury in patients with severe coronavirus disease 2019. Eur Heart J 2020;41:2070-9.
10. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017;38:500-7.
11. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, et al. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. JACC Cardiovasc Imaging 2019;12:681-9.
12. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, 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.
13. Lang RM, Addetia K, Miyoshi T, Kebed K, Blitz A, Schreckenberg M, et al. Use of machine learning to improve echocardiographic image interpretation workflow: a disruptive paradigm change. J Am Soc Echocardiogr 2021;34:443-5.
14. Narang A, Bae R, Hong H, Thomas Y, Surette S, Cadieu C, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol 2021;6:1-9.
15. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice. Circulation 2018;138:1623-35.
16. Ulloa Cerna AE, Jing L, Good CW, vanMaanen DP, Raghunath S, Suever JD, et al. Deep-learning-assisted analysis of echocardiographic

videos improves predictions of all-cause mortality. Nat Biomed Eng 2021; 5:546-54.
17. Drake DH, De Bonis M, Covella M, Agricola E, Zangrillo A, Zimmerman KG, et al. Echocardiography in pandemic: front-line perspective, expanding role of ultrasound, and ethics of resource allocation. J Am Soc Echocardiogr 2020;33:683-9.
18. Kirkpatrick JN, Mitchell C, Taub C, Kort S, Hung J, Swaminathan M. ASE statement on protection of patients and echocardiography service providers during the 2019 novel coronavirus outbreak: endorsed by the American College of Cardiology. J Am Soc Echocardiogr 2020;33: 648-53.
19. Karagodin I, Carvalho Singulane C, Woodward GM, Xie M, Tucay ES, Tude Rodrigues AC, et al. Echocardiographic correlates of in-hospital death in patients with acute COVID-19 infection: the World Alliance Societies of Echocardiography (WASE-COVID) study. J Am Soc Echocardiogr 2021;34:819-30.
20. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, 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.
21. Muraru D, Badano LP. Quantitative analysis of the left ventricle by echocardiography in daily practice: as simple as possible, but not simpler. J Am Soc Echocardiogr 2014;27:1025-8.
22. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, et al. Artificial intelligence in cardiovascular imaging: *JACC* state-of-the-art review. J Am Coll Cardiol 2019;73:1317-35.
23. Medvedofsky D, Mor-Avi V, Amzulescu M, Fernández-Golfín C, Hinojar R, Monaghan MJ, et al. Three-dimensional echocardiographic quantification of the left-heart chambers using an automated adaptive analytics algorithm: multicentre validation study. Eur Heart J Cardiovasc Imaging 2018;19:47-58.
24. Knackstedt C, Bekkers SC, Schummers G, Schreckenberg M, Muraru D, Badano LP, et al. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. J Am Coll Cardiol 2015;66:1456-66.
25. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018;20:65.
26. Tao Q, Yan W, Wang Y, Paiman EHM, Shamonin DP, Garg P, et al. Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 2019;290:81-8.
27. Asch FM, Yuriditsky E, Prakash SK, Roman MJ, Weinsaft JW, Weissman G, et al. The need for standardized methods for measuring the aorta: multimodality core lab experience from the GenTAC registry. JACC Cardiovasc Imaging 2016;9:219-26.
28. Farsalinos KE, Daraban AM, Ünlü S, Thomas JD, Badano LP, Voigt JU. Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the EACVI/ASE Inter-Vendor Comparison Study. J Am Soc Echocardiogr 2015;28:1171-81.e2.
29. Lim LJ, Tison GH, Delling FN. Artificial intelligence in cardiovascular imaging. Methodist Debakey Cardiovasc J 2020;16:138-45.
30. Quer G, Muse ED, Nikzad N, Topol EJ, Steinhubl SR. Augmenting diagnostic vision with AI. Lancet 2017;390:221.
31. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J 2019;40:2058-73.
32. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. J Am Coll Cardiol 2016;68:2287-95.
33. Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, et al. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. Circ Cardiovasc Imaging 2016;9:e004330.
34. Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med 2018;1:59.
35. Kawakami H, Wright L, Nolan M, Potter EL, Yang H, Marwick TH. Feasibility, reproducibility, and clinical implications of the novel fully automated assessment for global longitudinal strain. J Am Soc Echocardiogr 2021;34:136-45.e2.

ADDITIONAL WASE COVID INVESTIGATORS

Vince Ryan V. Munoz, MD (Philippine Heart Center, Quezon City, Philippines); Rafael Porto De Marchi, MD (Radiology Institute of the University of São Paulo Medical School, São Paulo, Brazil); Sergio M. Alday-Ramirez, PhD, and Consuelo Orihuela, MD (Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico); Anita Sadeghpour, MD, FASE (Rajaie Cardiovascular Medical and Center, Echocardiography Research Center, Iran University of Medical Science, Tehran, Iran); Jonathan Breeze, MD, and Amy Hoare (King's College Hospital, London, United Kingdom); Carlos Ixcanparij Rosales, MD (Centro Nacional 20 de Noviembre, ISSSTE, Mexico City, Mexico); Ariel Cohen, MD (Hôpitaux de l'Est Parisien St. Antoine-Tenon, Universite Pierre et Marie Curie, Paris, France); Martina Milani, MD, Ilaria Trolese, RDCS, Oriana Belli, MD, and Benedetta De Chiara, MD (Ospedale Niguarda, Milan, Italy); Michele Bellino, MD, and Giuseppe Iuliano, MD (University of Salerno, Salerno, Italy); and Yun Yang (Union Hospital, Tongji Medical College of HUST, Wuhan, China).

SUPPLEMENTARY APPENDIX

DESCRIPTION OF ECHOGO METHODS

The image processing pipeline used consists of two primary AI-based methods, a view classifier and an auto-contouring model, both composed of two-dimensional (2D) convolutional neural networks (CNNs).

All studies consisted of a set of 2D videos in Digital Imaging and Communications in Medicine format. CNN frameworks were developed in Python version 3.5 with Keras version 2.2.4 and TensorFlow version 1.13.

For the view classification model, a bespoke CNN was built using 10 convolutional layers that identifies apical two-chamber (A2C), apical 4CH (A4C), short-axis (SAX), and apical three-chamber (A3C) views acquired with and without contrast, respectively. Before view classification model training, images were processed using standard techniques to ensure homogenous and normalized image inputs. For the view classification model, the categorical cross entropy between the predicted view probability and manual label was used as loss function. For optimization, the RMSprop optimizer was used with an initial learning rate of 1e-4 and a decay rate of 1e-6 to allow time inverse decay of learning rate. We ran our training for 50 epochs or about 165,000 iterations, and the one with the highest validation accuracy was selected as the best model. Real-time data augmentation was applied during training to improve model generalizability.

A U-net-based CNN segmentation frameworks (see Supplemental Figure X for the model architecture) was developed to contour the LV endocardium in A2C and A4C studies. End-diastolic (ED) and end-systolic (ES) image labels were used to train the CNNs initially contoured manually by British Society of Echocardiography (BSE)–accredited echocardiographers.

Raw images were processed and fed into the modified U-net CNN framework along with their corresponding contours (which were converted to binary masks so the model could be trained as a segmentation problem). Real-time data augmentation was performed during training to increase the accuracy of the model. The efficacy of the network's segmentation performance was assessed using the Sørensen-Dice coefficient (DC), and this was used as the training loss function. The optimization was performed using the Adam optimizer, with the initial learning rate set to 1e-4, and run for 400 epochs. The model with the lowest validation loss was selected for use.

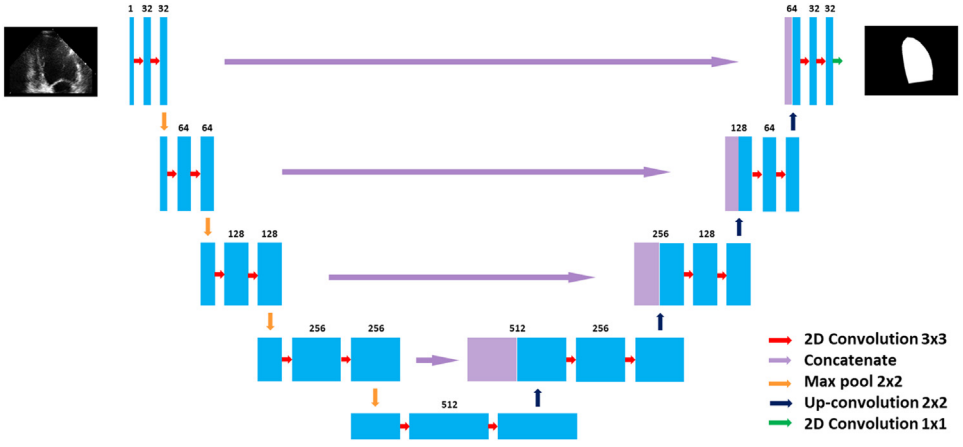
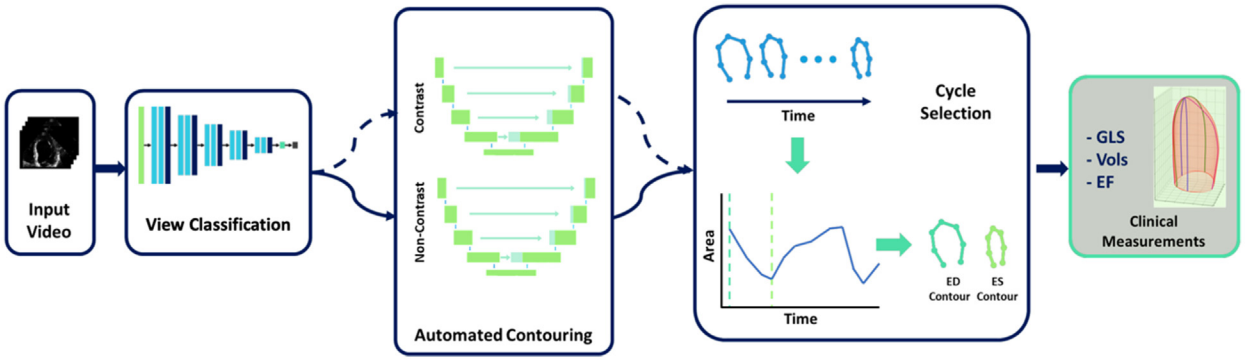


Fig. 1 Modified U-net architecture used for auto-contouring. The feature depth of the layers is substantially smaller than those used in the original U-net because the convolutions were set to have half the number of filters.

DEVELOPMENT OF VIEW CLASSIFICATION



Model development and validation were based on a data set of clinical information and images extracted from several multicenter studies. Images had been acquired from hospitals with a range of sizes, types of operators, and ultrasound vendor equipment representative of "real world" echocardiography. Data were accessed with participants' informed consent, and ethical approval for the study was obtained from Health Research Authority National Research Ethics Service Committee South Central – Berkshire (Integrated Research Application System reference 14/SC/1437). For model training, a set of images were identified that included A4C and A2C views, had endocardial visualization in ≥14 of 16 segments in available images (on the basis of consensus review by three BSE-accredited cardiac physiologists), and had ED and ES frames with a minimum of four frames between end-diastole and end-systole. The echocardiograms were used to develop a series of sequential image processing algorithms. All studies consisted of a set of 2D videos in Digital Imaging and Communications in Medicine format. CNN frameworks were developed using Python version 3.5 with Keras version 2.2.4 and TensorFlow version 1.13. For the view classification model, a bespoke CNN was built using 10 convolutional layers that identifies A2C, A4C, SAX, and A3C views acquired with and without contrast, respectively. Training data comprised 1,250 2D echocardiograms from 1,014 subjects.

Development of View Classification

Before view classification model training, images were processed using standard techniques to ensure homogenous and normalized image inputs to the training pipeline. Before processing, the data were split into 90% for model training (211,958 image frames) and 10% used as a testing data set (23,946 frames). Image frames from different subjects were separated out entirely among the training and testing data sets. Furthermore, a separate validation data set of 240 studies was acquired from the ultrasound testing data and used as an independent test data set (39,401 frames). Before independent validation of the entire sequential AI pipeline in the quarantined validation datasets, image frames from different subjects were separated out entirely among the training and testing data sets. Overall accuracy of 95% was achieved for the CNN view classifier used to identify eight echocardiographic views.

View classifier final performance matrix.

Development of LV Segmentation

Following view classification, a U-net-based CNN segmentation framework was developed to contour the LV endocardium in the A2C and A4C views. ED and ES image labels were used to train the CNNs images initially contoured manually by three BSE-accredited echocardiographers. The model was trained from data comprising 5,692 frames. The data sets were split into 80% training and 20% testing data sets for training the CNNs. Raw images were processed and fed into the modified U-net CNN framework. The CNNs produced contours that were able to track the endocardial walls smoothly through time. The efficacy of the network's segmentation performance was assessed using the DC. An algorithm was established from the image clips and LV contours to identify the cardiac cycle and ED and ES frames, comprising assessment of contour areas and R-wave triggers. Where the R-wave trigger was not available, the heart rate was inferred from the image LV segmentation. Heart rate extraction required the reduction of image dimensionality, followed by signal period extraction, using dimensionality reduction methodologies.

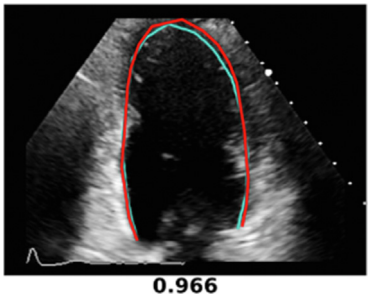
Performance of Auto-View Classification

Using an unseen data set of 240 studies (23,946 frames), overall accuracy of 95% was achieved for the CNN view classifier used to identify and label eight echocardiographic views. Classification accuracy of noncontrast views (A2C, A3C, A4C, and SAX) exceeded 97%, while the accuracy of contrast A3C, A4C, and SAX view classification

was >93%, and that of contrast A2C views was 85%. Differentiation between contrast and noncontrast views was 100%.

Performance of LV Contouring and Segmentation

For the auto-contouring model, a testing data set of 436 image frames were contoured by the model, which demonstrated high concordance with manual contours from three BSE-accredited echocardiographers, as measured by the DC. The model achieved a mean contrast A2C and A4C DC of 93.25%, exceeding that seen in other works. The auto-contouring model performed consistently well on data captured using multiple ultrasound vendors and models and did not exhibit significant declines in DC scores across the range of image quality considered clinically acceptable for stress echocardiography. For LV segmentation, included images met the following requirements: at least one of LV-focused A4C and standard A2C views present and endocardial visualization in ≥14 of 16 segments in available images by consensus of three BSE-accredited cardiac physiologists. A U-net-based CNN segmentation framework was developed to contour the LV endocardium in both the A2C and A4C views. ED and ES image labels were used to train the U-net-based CNN initially contoured manually by three BSE-accredited echocardiographers. The images comprised 5,692 and 2,182 frames for A2C and A4C views, respectively. Both data sets were split into 80% training and 20% testing data sets. Raw images were resized, filtered, normalized, and applied to the modified U-net CNN framework. The CNN produced contours that were able to track the endocardial walls smoothly through time. The efficacy of the network's segmentation performance was assessed using the DC. High concordance with manual contours from three BSE-accredited echocardiographers, as measured by the DC was achieved for the LV segmentation model. The model achieved a mean A2C and A4C DC of 92.47%.



Example of an LV AI auto and manual tracing.

Cycle and Frame Selection: A rule-based algorithmic method was established from the image clips to identify the cardiac cycle and ED and ES frames, comprising assessment of maximum and minimum LV segmentation areas in combination with the R-wave triggers or R-wave inference when the R-wave was not available. Frame selection was evaluated as part of the final AI pipeline independent validation.

Quantification of Measures: For LS, Euclidean distances between contour points was calculated, moving sequentially from the starting basal point to the end basal point. The sum of the individual Euclidean distance values is calculated over the entire contour. End-diastole (base length) and end-systole (shortened length) were used to calculate longitudinal strain as follows:

Longitudinal strain (%) = [(shortened length − base length)/(base length)] ×100.

For LVEF, the volume of the left ventricle is first calculated using Simpson's biplane method using the following formula:

$$V = \pi \times \delta \sum_{i=1}^{n_d} a_i b_i,$$

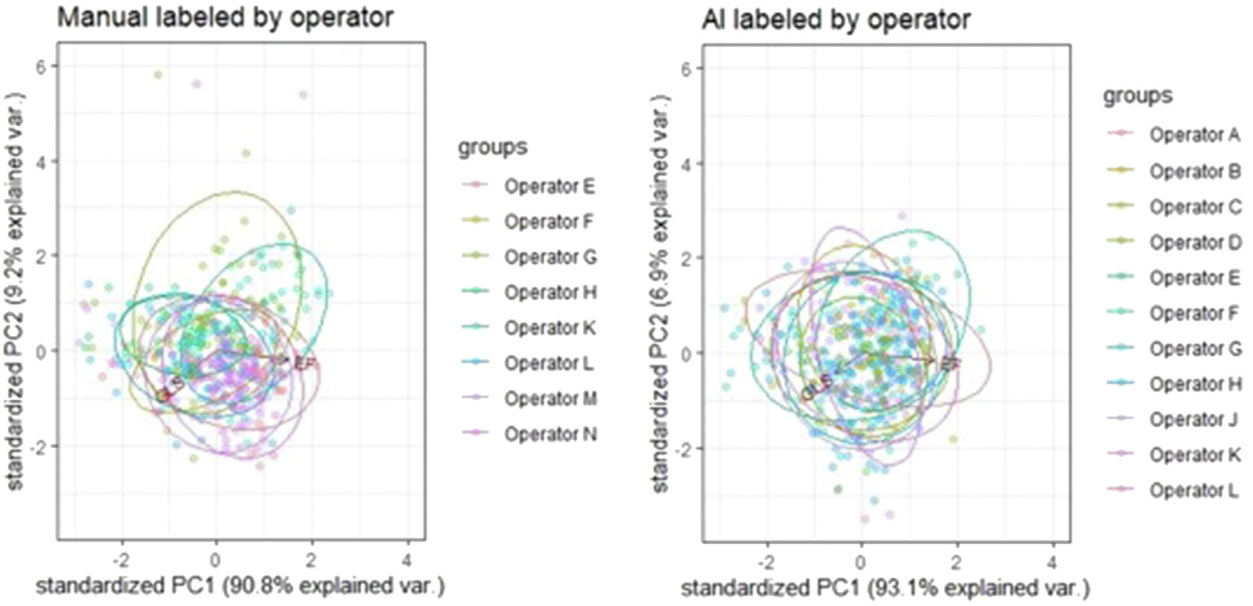
The contours are assumed to be directly perpendicular to one another and to meet at the apex as per the standard Simpson biplane method. The ejection fraction is calculated using the following formula:

$$EF = \frac{(V^D - V^S)}{V^D} \times 100$$

Accuracy Assessment during AI Development: To assess the accuracy and precision of the complete AI pipeline (comprising the view classification model, LV automated contouring model, and ED and ES frame selection algorithm), a quarantined data set of echocardiograms were processed through the AI pipeline and TomTec software by five qualified observers (three echocardiographers and two cardiologists) to derive global measures of LV function comprising ED and ES volumes, LVEF, and LVLS. Accuracy was defined as the mean value produced across observers during echocardiographic examination. As each measurement was made by a different independent observer, mean values (AI vs TomTec) as an overall assessment of performance was appropriate using standard Bland-Altman analysis and Deming regressions, because repeated measures were not observed in this instance. No observable differences in accuracies were noted between ultrasound vendors, with mean ES volume, ED volume, LVLS, and LVEF root mean square error values of 7.8 mL, 9.9 mL, 2.6%, and 5.3%, respectively for Philips models and 8.0 mL, 12.2 mL, 2.56%, and 4.75%, respectively for GE models. For LVLS, the overall mean bias was shown to be 0.19%, and root mean square error was 2.89%, consistent with literature reports of intervendor comparisons. For LVEF, the overall mean bias was −1.18%, with a root mean square error of 5.02%.

Principal-Component Analysis Methodology

Principal-component analysis (PCA) is a multivariate statistical method frequently used in exploratory data analysis. Although PCA is traditionally used as a dimensionality reduction method, by projecting each data point (each patient) onto the first two principal components, here we define the first principal component as a direction that maximizes the total variance of the projected data (patient read), on the basis of the input variables (LVEF and LVLS). In this manner, PCA plots the projected data on the basis of the variance rather than the dimension reduction. In effect, the closer the data points are together, the less variance can be thought to be between them. If unbiased, the projection of the data should be independent of the operator who produced the measurement. However, if bias is observed, groupings by operator will be observed, strongly suggesting operator influence on the observed LVEF and LVLS values. As such we see, visually, a higher degree of operator associated grouping for LVEF and LVLS values in manual reads than compared with the grouping observed for the AI reads, implying a greater operator-associated bias in the manual reads. The overall wider spread of datapoints in the manual reads implies a greater variance, while on AI reads this operator-associated bias is not observed to the same degree, with a tighter cluster of all data points, which implies less overall variance for this dataset.



Supplemental Figure 1 Operator influence on the variability in LVEF and LVLS in manual and AI measurements was visualized using PCA. PCA eigenvalues were calculated on the basis of LVLS and LVEF values from manual and AI contouring separately. Each data point was subsequently labeled according to the individual operator in order to investigate whether operator-based clustering was present. PCA visualizing the summary variability information contained in the data set described by LVEF and LVLS, colored by operator (each operator is named with a letter and represented with a color). Each data point identifies a TTE study, and each TTE study is colored by the operator who performed the contouring. *(Left)* PCA of manual contouring clusters the TTE studies by operator, identifying the operator as a possible confounder. *(Right)* No clustering (i.e., good grouping of points) is present in the PCA on AI-contoured TTE studies.

Supplemental Table 1 Univariable logistic regression to outcomes by reading round

	In-hospital death		Death at follow-up	
Parameter	Odds ratio [95% CI]	*P*	Odds ratio [95% CI]	*P*
LVEF manual				
Round 1	0.988 (0.970-1.006)	.104	0.995 (0.978-1.012)	.568
Round 2	0.986 (0.969-1.003)	.104	0.988 (0.972-1.004)	.137
Round 3	0.987 (0.970-1.004)	.119	0.987 (0.972-1.003)	.108
LVEF AI				
Round 1	0.971 (0.953-0.989)	.002	0.975 (0.958-0.992)	.005
Round 2	0.976 (0.958-0.995)	.013	0.985 (0.967-1.004)	.109
LVLS manual				
Round 1	1.012 (0.976-1.050)	.521	1.011 (0.979-1.045)	.510
Round 2	1.002 (0.960-1.048)	.908	1.007 (0.976-1.040)	.658
Round 3	1.045 (1.009-1.085)	.017	1.038 (1.004-1.076)	.033
LVLS AI				
Round 1	1.080 (1.034-1.130)	<.001	1.057 (1.017-1.101)	.006
Round 2	1.068 (1.020-1.119)	.005	1.049 (1.006-1.096)	.025

Supplemental Table 2 Pairwise Pearson correlation (*r*) matrix to clinical measures for AI and manual reads

	LVLS	LVEF	BNP	CRP	SBP	DBP
Manual reads						
LVLS	1.000	−0.735	0.499	0.146	−0.113	−0.082
LVEF	−0.735	1.000	−0.517	−0.102	0.081	0.043
AI reads						
LVLS	1.000	−0.744	0.336	0.235	−0.176	−0.149
LVEF	−0.744	1.000	−0.467	−0.219	0.185	0.199

CRP, C-reactive protein; *DBP*, diastolic blood pressure; *SBP*, systolic blood pressure.
Only those significant after Bonferroni correction are displayed in the network.

Supplemental Table 3 Cox proportional-hazards regression against in-hospital outcomes across AI and manual reads

	Outcome: in-hospital mortality	
	Hazard ratio (95% CI)	*P*
Echocardiographic parameters (continuous)		
LVEF manual	0.988 (0.973-1.003)	.110
LVEF AI	0.979 (0.963-0.995)	.011
LVLS manual	1.031 (0.996-1.066)	.080
LVLS AI	1.046 (1.005-1.089)	.028
LVESV manual (log_2)	1.026 (0.777-1.355)	.855
LVESV AI (log_2)	1.146 (0.849-1.547)	.373
LVEDV manual (log_2)	0.999 (0.654-1.525)	.995
LVEDV AI (log_2)	0.972 (0.623-1.518)	.901
Echocardiographic parameters (categorical)		
LVEF manual (reference: <60%)		
>60%	0.833 (0.533-1.30	.422
LVEF AI (reference: <60%)		
>60%	0.571 (0.365-0.894)	.014
LVLS manual (reference: <−16%)		
>−16%	1.701 (1.075-2.692)	.023
LVLS AI (reference: <−16%)		
>−16%	1.721 (1.100-2.691)	.017
Significant clinical parameters		
Age	1.026 (1.010-1.044)	.002
Outcome		
ICU	3.907 (2.304-6.625)	<.001
Ventilator	4.512 (2.853-7.136)	<.001
Hemodynamic support	3.503 (2.221-5.526)	<.001
Previous conditions		
Lung disease	1.775 (1.033-3.048)	.038
Heart disease	1.638 (0.991-2.707)	.054
Biomarkers		
CRP (reference: normal)		
Borderline	3.546 (0.222-56.70)	.371
Abnormal	12.157 (1.688-87.540)	.013
BNP (reference: normal)		
Borderline	2.245 (0.709-7.108)	.169
Abnormal	3.366 (1.444-7.846)	.005
DBP	0.967 (0.946-0.989)	.003

CRP, C-reactive protein; *DBP*, diastolic blood pressure; *LVEDV*, LV end-diastolic volume; *LVESV*, LV end-systolic volume.