Deep Learning for Improved Precision and Reproducibility of Left Ventricular Strain in Echocardiography: A Test-Retest Study

Ivar M. Salte, MD, Andreas Østvik, MSc, PhD, Sindre H. Olaisen, MD, Sigve Karlsen, MD, Thomas Dahlslett, MD, Erik Smistad, MSc, PhD, Torfinn K. Eriksen-Volnes, MD, Harald Brunvand, MD, PhD, Kristina H. Haugaa, MD, PhD, Thor Edvardsen, MD, PhD, Håvard Dalen, MD, PhD, Lasse Lovstakken, MSc, PhD, and Bjørnar Grenne, MD, PhD, *Trondheim, Kristiansand, Oslo, Arendal, Levanger, Norway; and Stockholm, Sweden*

Aims: Assessment of left ventricular (LV) function by echocardiography is hampered by modest test-retest reproducibility. A novel artificial intelligence (AI) method based on deep learning provides fully automated measurements of LV global longitudinal strain (GLS) and may improve the clinical utility of echocardiography by reducing user-related variability. The aim of this study was to assess within-patient test-retest reproducibility of LV GLS measured by the novel AI method in repeated echocardiograms recorded by different echocardiographers and to compare the results to manual measurements.

Methods: Two test-retest data sets (n = 40 and n = 32) were obtained at separate centers. Repeated recordings were acquired in immediate succession by 2 different echocardiographers at each center. For each data set, 4 readers measured GLS in both recordings using a semiautomatic method to construct test-retest interreader and intrareader scenarios. Agreement, mean absolute difference, and minimal detectable change (MDC) were compared to analyses by AI. In a subset of 10 patients, beat-to-beat variability in 3 cardiac cycles was assessed by 2 readers and AI.

Results: Test-retest variability was lower with AI compared with interreader scenarios (data set I: MDC = 3.7 vs 5.5, mean absolute difference = 1.4 vs 2.1, respectively; data set II: MDC = 3.9 vs 5.2, mean absolute difference = 1.6 vs 1.9, respectively; all $P < .05$). There was bias in GLS measurements in 13 of 24 test-retest interreader scenarios (largest bias, 3.2 strain units). In contrast, there was no bias in measurements by AI. Beat-to-beat MDCs were 1,5, 2.1, and 2.3 for AI and the 2 readers, respectively. Processing time for analyses of GLS by the AI method was 7.9 $\pm$ 2.8 seconds.

Conclusion: A fast AI method for automated measurements of LV GLS reduced test-retest variability and removed bias between readers in both test-retest data sets. By improving the precision and reproducibility, AI may increase the clinical utility of echocardiography. (J Am Soc Echocardiogr 2023;36:788-99.)

Keywords: Left ventricular function, Echocardiography, Strain, Reproducibility, Artificial intelligence

From the Department of Medicine, Hospital of Southern Norway, Kristiansand, Norway (I.M.S.); Faculty of Medicine, University of Oslo, Oslo, Norway (I.M.S., S.K., K.H.H., T.E.); Centre for Innovative Ultrasound Solutions and Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway (A.O., S.H.O., E.S., T.K.E.V., H.D., L.L., B.G.); Medical Image Analysis, Health Research, SINTEF Digital, Trondheim, Norway (A.O., E.S.); Department of Medicine, Hospital of Southern Norway, Arendal, Norway (S.K., T.D., H.B.); ProCardio Center for Innovation, Department of Cardiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway (I.M.S., K.H.H., T.E.); Faculty of Medicine, Karolinska Institutet and Cardiovascular Division, Karolinska University Hospital, Stockholm, Sweden (K.H.H.); Clinic of Cardiology, St. Olavs University Hospital, Trondheim, Norway (T.K.E.V., H.D., B.G.); and Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway (H.D.).

Drs. Salte and Østvik contributed equally to this work.

This work was supported by the Centre for Innovative Ultrasound Solutions (a Norwegian Research Council center for research-based innovation, project no. 237887), by the Norwegian Health Association, the Central Norway regional health authority, South-Eastern Norway regional health authority, the national program for clinical therapy research (project no. 2017207), and ProCardio Center for Innovation (no. 309762) from the Norwegian Research council. The HUNT Echocardiography study was funded by the Liaison Committee for Education, Research and Innovation in Central Norway and grants from the Simon Fougner Hartmann Family Fund, Denmark.

Conflicts of Interest: None.

Reprint requests: Bjørnar Grenne, MD, PhD, FESC, FEACVI, Clinic of Cardiology, St. Olavs University Hospital, Postbox 3250, Torgarden NO-7006 Trondheim, Norway; Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway (E-mail: *bjornar.grenne@ntnu.no*).

0894-7317

https://doi.org/10.1016/j.echo.2023.02.017

Abbreviations

AI = Artificial intelligence

ASE = American Society of Echocardiography

GLS = Global longitudinal strain

ICC = Intraclass correlation coefficient

LV = Left ventricle, ventricular

LVEF = Left ventricular ejection fraction

MDC = Minimal detectable change

ROI = Region of interest

INTRODUCTION

Reliable test-retest reproducibility is critical for the utility of diagnostic tests, although rarely assessed or discussed in echocardiographic studies. Suboptimal test-retest reproducibility hampers traditional quantification of left ventricular (LV) ejection fraction (LVEF), which is crucial in everyday decision-making for diagnosis, follow-up, prognostic evaluation, and treatment in large patient groups.[1-5] Left ventricular global longitudinal strain (GLS) has been introduced as a parameter for LV function that could outperform LVEF, in terms of both reproducibility and prognostic value.[1] However, the present semiautomatic methods for analyses of LV GLS are still limited by reader dependency that introduces measurement variability and adds to the time-consuming process of analyzing echocardiographic images.[6] Thus, measurement of LV GLS is underused in everyday clinical practice. To overcome these challenges, there is need for a fast, feasible, and more reproducible method to gain diagnostic and clinical benefits.

Deep learning, one of the most recent advancements in machine learning and a key component in artificial intelligence (AI), now enables computer algorithms to learn from annotated images without prior feature extraction. The field of deep learning can lead to a paradigm shift in cardiac imaging by changing the echocardiographic workflow.[7] By reducing time-consuming manual measurements and the variability related to image interpretation, both the reproducibility, efficiency, and efficacy of echocardiography may be improved. Automated AI-based measurements of LV GLS could also improve diagnostic and prognostic accuracy.[8]

Recently, a fully automated deep learning–based AI method was shown to provide high feasibility and accuracy for measurements of LV GLS.[9] This is the first AI-based GLS software to include a deep-learning network specifically trained to perform the motion estimation task, which could improve tracking accuracy compared with contour tracing or traditional block- and feature-matching algorithms.[10] However, although a fully automated deep-learning algorithm reproduces the same result every time when applied to the exact same images, repeated echocardiographic recordings always introduce image differences due to variations in probe positioning, angulation, and tilt, as well as the patient's position, breathing, and heart rate. Therefore, it is of great importance to quantify how an automated AI method influences measurement agreement when analyzing repeated echocardiograms within patients. Such knowledge is lacking for automated measurements of LV GLS.

Thus, we aimed to study the test-retest reproducibility of LV GLS measured by the fully automated AI method compared with an established semiautomatic method when analyzing within-patients repeated echocardiographic recordings acquired by different echocardiographers. There is no easily obtainable gold standard for true LV GLS, and the purpose of our study was not to assess the accuracy of the established or the novel method.

METHODS

Study Design Overview

We performed a reproducibility study in 2 data sets of test-retest echocardiographic recordings from 2 independent academic centers in Norway (Graphical Abstract). The study was designed to simulate a realistic clinical test-retest situation where LV GLS was measured in images from 2 separate recording sessions in each patient, acquired by 2 different echocardiographers. To minimize the variability caused by differences in physiological conditions, the test-retest recording sessions were acquired in immediate succession. Each of the test-retest data sets was recorded by 2 different echocardiographers for each institution, in total 4 different echocardiographers. Both recordings in each patient were analyzed by a total of 4 readers, that is, the 2 who recorded the echocardiograms and 2 not participating in the image recording process. The latter 2 readers analyzed the data sets from both institutions. All readers measured LV GLS using a semiautomatic reference method. Thus, a total of 6 readers participated in the study, labeled by letters from A to F. Supplemental Online Table 1 lists each reader's medical position and experience in transthoracic echocardiography. Readers A and B analyzed both data sets from each institution. Readers C and D were unique for data set I, whereas readers E and F were unique for data set II. For each data set, this allowed for the construction of 12 unique test-retest scenarios where the 2 recordings were analyzed by different readers (test-retest interreader scenarios) and 4 scenarios where the 2 recordings were analyzed by the same reader (test-retest intrareader scenarios). All measurements were performed blinded to clinical data and the results of other readings. Finally, the repeated recordings in each patient were analyzed by the AI method. The agreement of the repeated-recording test-retest inter- and intrareader scenarios was assessed and compared to the results when both recordings were analyzed by the AI method.

Material

Data set I was from a cohort of patients with a history of hospitalization due to suspected acute coronary syndrome and was collected at Sørlandet Hospital Arendal, Norway. Data set II was collected as part of the Trøndelag Health Study (the HUNT Study), a cross-sectional health study in central Norway. Both data sets included repeated echocardiographic recordings in random samples of the study populations, performed to investigate measurement variability in echocardiography. Complete test-retest echocardiograms with cine-loops from the 3 standard apical views were available for 40 and 32 subjects, respectively. There was no selection based on cardiac disease or image quality, and thus the 2 data sets contained echocardiographic recordings with a wide range of cardiac function and image quality.

Echocardiographic acquisitions were performed with GE Vivid 7 (data set I) and GE Vivid E95 (data set II), both from GE Vingmed Ultrasound. Acquisitions were performed in accordance with European Association of Cardiovascular Imaging and American Society of Echocardiography (ASE) recommendations.[11,12]

Image quality was visually assessed per segment in a standard 18-segment LV model. Each segment was scored as missing if outside the image sector or if the myocardium was indistinguishable from surrounding structures due to artifacts. Examinations were classified as good quality if no segments were missing from any of the 3 apical views, fair quality if 1 to 2 segments were missing, and poor quality if >2 segments were missing.

HIGHLIGHTS

- Deep-learning AI provides efficient automated GLS measurements in echocardiograms.
- Deep-learning AI produces consistent GLS measurements in repeated echocardiograms.
- Automated GLS measurements using deep learning improve test-retest reproducibility.

The study was approved by the Regional Committee for Medical and Health Research Ethics (REC IDs 53,266 and 13,083) and was conducted in compliance with the ethical principles of the Declaration of Helsinki.

Global Longitudinal Strain Measured by the Semiautomatic Method

Semiautomatic LV GLS was analyzed with a commercially available and widely used speckle-tracking software (2DS, EchoPAC SWO ver. 203, GE Ultrasound). This reference method is one of the most well-studied applications for strain measurements, and one of the few being validated using both sonomicrometry and cardiac magnetic resonance imaging.[13] Measurements were performed as recommended by the European Association of Cardiovascular Imaging and ASE.[12] End diastole was defined by the semiautomatic software and only corrected if needed by visual assessment. End systole was identified manually by the aortic valve closure. The readers identified endocardial and epicardial borders by visual assessment and manually corrected the default region of interest (ROI) proposed by the speckle-tracking software if needed. Manual ROI adjustment was required in most patients. Software-specific default values of spatial and temporal smoothing were used, and automatic drift compensation was applied by default. Examinations were rejected if >2 adjacent segments of a single view were missing. Left ventricular GLS was calculated as the average from the 3 standard apical views. A representative single beat was selected and analyzed from each of the standard apical views. In addition, 2 readers analyzed 3 consecutive beats per recording in a random subset of 10 patients to assess beat-to-beat variability.

To quantify possible differences in manual adjustments of the ROI initiation between readers, the end-diastolic ROI centerline length and ventricular length were calculated on the basis of ROI centerline positional data provided by the semiautomatic software, which were available for reader A (ASE level II, experience: >300 strain analyses) and reader B (ASE level II, experience: >50 strain analyses).

Global Longitudinal Strain Measured by the AI Method

An in-house-developed AI method based on deep learning was used to perform automated image analyses and measurements of LV GLS (Figure 1). The AI method utilized artificial neural networks to perform key tasks such as image view classification, cardiac event timing, image segmentation, and motion estimation. The components of the AI method were trained using different databases and training strategies. The view classification was trained on approximately 250 patients, the event timing model on 500 patients, and the segmentation model on more than 600 patients. The motion estimation method was first pretrained on roughly 50,000 image pairs of synthetic data of different moving objects rendered on random backgrounds in addition to sequences from an animation movie. After this, transfer learning was conducted on 105 video sequences, or roughly 3,000 image pairs of simulated ultrasound data with ground truth motion derived from a biomechanical model. Finally, 100 recordings of real patient data were used for fine-tuning of the model. For this step, image quality was first assured by an expert, followed by extensive augmentation based on ROI initiation and motion tracking made by the semiautomatic speckle-tracking method. All databases included patients with large variation in LV morphology and function. To measure LV GLS, reference points were seeded along the centerline of the ROI defined by the segmentation network in the frame classified as end diastole by the timing network. The line drawn through these points constituted the length of the myocardium at baseline. The positions of the reference points were updated by the motion estimation network per frame through the cardiac cycle. In contrast to traditional methods for motion estimation in echocardiography, this novel approach applies a deep neural network to estimate myocardial motion by using state-of-the-art, learning-based optical flow mapping tailored for ultrasound images, where the myocardial displacement is estimated between successive frames.[12] Additional details regarding the AI method have been described elsewhere.[9,10]

Left ventricular GLS was calculated as the Lagrangian peak negative strain. Similar to the reference method, peak strain was calculated for all 3 standard apical views, and the reported LV GLS was calculated as the average of these 3 values. The AI method measured and reported LV GLS based on the single middle beat of the 3 cycles of each recorded view (1 beat) and as the beat-to-beat average of all 3 cycles (3-cycle beat-to-beat average).

Beat-to-Beat Variability

Beat-to-beat variability was studied by randomly selecting 10 patients from data set I. Two blinded readers (readers A and B) analyzed GLS in 3 consecutive cardiac cycles for each of the 3 apical views in both the first and second echocardiographic recordings. The exact same cine-loops and cycles were analyzed by both readers. This resulted in 60 cine-loops of 3 consecutive beats analyzed by both readers and a total of 360 reference measurements. The beat-to-beat variability by the 2 readers was compared to the results by the AI method.

Statistics

As data were normally distributed, continuous variables are presented as mean ± SD. Categorical variables are presented as numbers and percentages. Bland-Altman analyses were performed to assess test-retest measurement variability. Bias and limits of agreement were calculated for each test-retest scenario.[14] Measurement reproducibility was quantified by estimating the standard error of measurement (SE_M) calculated as the root mean squared average of within-patient SDs. We calculated the minimal detectable change (MDC) as $1.96 \times \sqrt{2}$ times the SE_M. In beat-to-beat assessments, the SE_M and MDC were calculated using the within-recording SDs. The coefficient of variation was calculated as SE_M divided by the mean of all measurement pairs multiplied by 100. Intraclass correlation coefficients (ICCs) were calculated using a 2-way mixed-effect absolute agreement model. A 2-sided paired *t* test was used to test whether the average within-patient SDs of 2 scenarios were statistically different. The difference between AI and the average interobserver and intraobserver scenarios was calculated for mean absolute difference, SE_M, MDC, coefficient of variation, and ICC. The jackknife technique was used to calculate the SE of the difference estimates

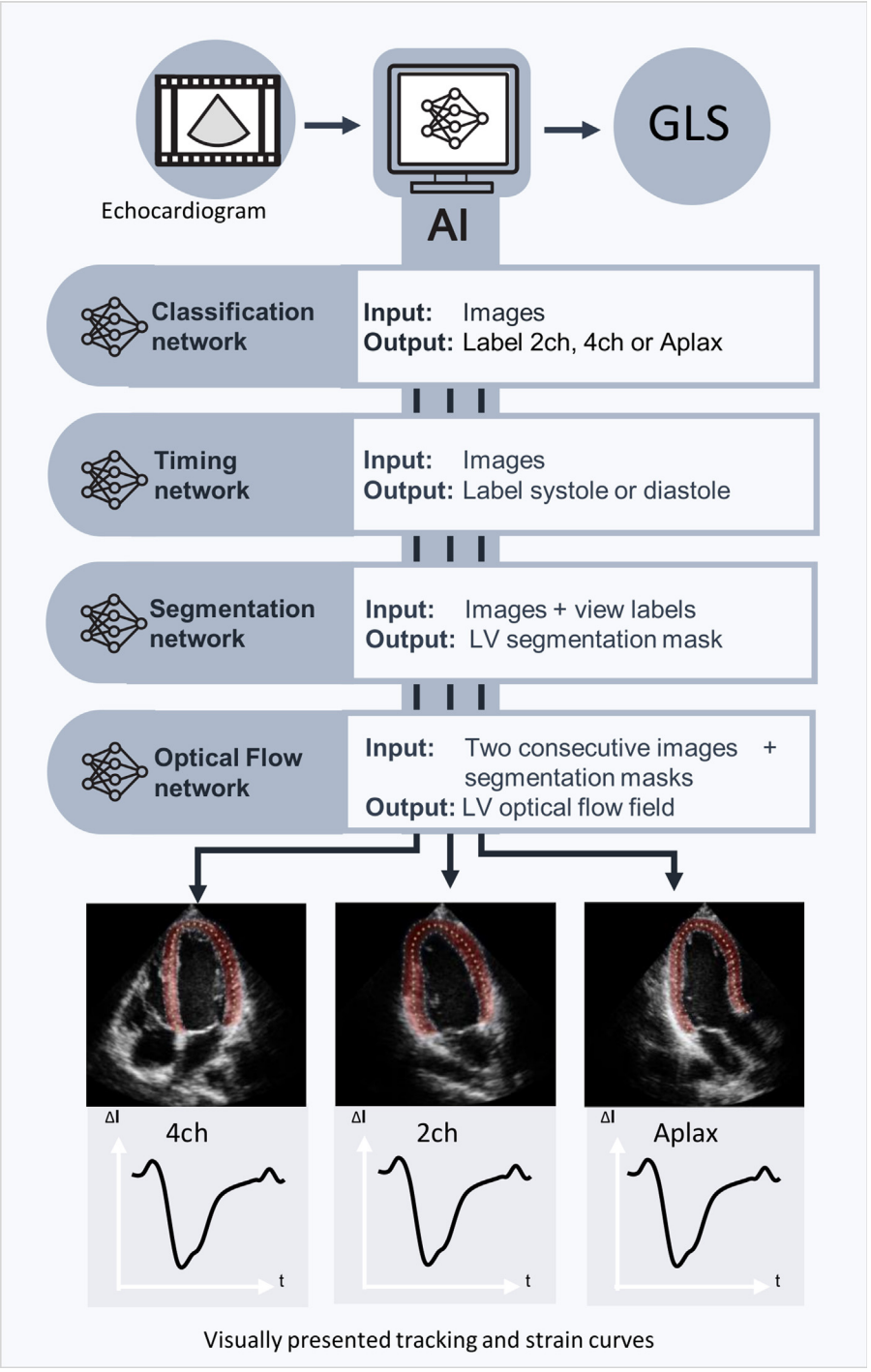

Figure 1 Schematic illustration of our in-house-developed AI method for automated measurements of LV GLS. The input was echocardiographic studies containing 4-chamber (4ch), 2-chamber (2ch), and apical long-axis views (Aplax). Four deep-learning networks were used for the key tasks of view classification, timing of cardiac events, image segmentation, and motion estimation. To measure LV GLS, the current view was defined by the view classification network, the end-diastolic frame was detected by the timing network, and a line was drawn through points seeded along the centerline of the myocardial segmentation mask. The position of these seeded points and the resulting centerline of the myocardium were updated through the cardiac cycle by the flow fields produced by the motion estimation network. Lagrangian peak negative strain was measured in the 3 apical views, and the average GLS was reported.

and a Z test was used to test whether the differences were significantly different from 0. $P < .05$ was considered statistically significant.

All statistical analyses were performed using Python 3.7.4 (Python Software Foundation) code based on open-source statistical Python packages (SciPy 1.5.4, Pingouin 0.5.3, and Statsmodels 0.12.1). Exact 95% CI of the limits of agreement were calculated using code based on the method proposed by Shieh.[15]

RESULTS

Demographic characteristics of the 2 populations are summarized in Table 1. Patients in data set I were older and had slightly lower LVEF and more comorbidity compared with patients in data set II. None of the patients were excluded based on image quality. The AI method succeeded in classifying the correct view in 96% (231/240) of the recordings in data set I and 97% (187/192) of the recordings in data set II. Further, the AI method correctly classified cardiac events (end diastole, systole, and end systole) in 99% (238/240) of recordings in data set I and 97% (187/192) of recordings in data set II. Image segmentation, estimation of cardiac motion, and measurement of LV GLS were possible in all examinations when the correct view and timing of events were verified. Total processing time for LV GLS per patient was 7.9 ± 2.8 seconds.

Data Set I Test-Retest Reproducibility

In data set I, the mean LV GLS measured by the 4 readers using the semiautomatic reference method ranged from −17.2% ± 3.0% to −20.1% ± 3.2%, whereas LV GLS measured by the AI method was −16.0% ± 2.4%. The average MDCs, mean absolute differences, SE_M, coefficients of variation, and ICCs of the test-retest scenarios are presented in Table 2. Compared with the mean of the interreader scenarios, use of AI reduced MDC (3.7 vs 5.5, respectively, $P < .05$).

When LV GLSs in the 2 recordings were analyzed by different readers (interreader scenarios), a significant bias between readers was observed in 9 of 12 scenarios, with a largest absolute bias of 3.2 strain units (Figure 2). When LV GLSs in the 2 recordings were analyzed by the same reader (intrareader scenarios), a significant bias of 0.8 strain units was found in 1 of 4 scenarios (Figure 3). Using AI for measurement of LV GLS in both recordings (AI scenario) resulted in no significant bias.

Data Set II Test-Retest Reproducibility

In data set II, mean LV GLS measured by the 4 readers using the semiautomatic reference method ranged from −17.7% ± 2.6% to −19.2%% ± 2.7%%, whereas LV GLS measured by AI was −16.8% ± 2.7%. The average MDCs, mean absolute differences, SE_M, coefficients of variation, and ICCs of the test-retest scenarios are presented in Table 2. Compared with the mean of the interreader scenarios, use of AI reduced MDC (3.9 vs 5.2, respectively, $P < .05$).

When LV GLSs in the 2 recordings were analyzed by different readers (interreader scenarios), a significant bias between readers was observed in 4 of 12 scenarios, with the largest absolute bias of 1.6 strain units (Figure 4). When LV GLS in the 2 recordings was analyzed by the same reader (intrareader scenarios), there was no significant bias observed in any of the scenarios (Figure 5). Similarly, using AI for measurement of LV GLS in both recordings (AI scenario) resulted in no significant bias.

Beat-to-Beat Reproducibility Substudy

Beat-to-beat reproducibility of LV GLS in 3 consecutive cardiac cycles was improved when measurements were performed by AI compared with conventional semiautomatic measurements by the 2 readers ($SE_M = 0.55$, 0.75, and 0.84 for AI, reader A, and reader B, respectively, $P < .05$). Correspondingly, the MDC was lower for the AI method compared with the 2 readers (MDC = 1.5, 2.0, and 2.3 for AI, reader A, and reader B, respectively, $P < .05$).

Influence of Image Quality on Test-Retest Variability

There was a trend toward lower mean absolute difference with better image quality. Mean absolute difference (SD) strain (%) for AI measurements and the intraobserver scenarios were 2.0 (1.3) and 1.9 (1.2) and in recordings graded as having poor image quality. Correspondingly, in recordings with good image quality, the mean absolute difference was approximately 40% lower, with a mean absolute difference (SD) strain (%) of 1.3 (0.9) and 1.2 (0.7), respectively, with overlapping CIs according to image quality and between methods (Supplemental Online Figure 1).

DISCUSSION

This is the first study to demonstrate that measurements of LV GLS using a fully automated AI method based on deep learning improves within-patient test-retest reproducibility in echocardiography. The test-retest reproducibility of AI-based measurements was favorable compared to interreader scenarios and comparable to the intrareader scenarios. In repeated echocardiographic examinations performed by different echocardiographers, the bias observed in the interobserver scenarios, representing systematic between-operator differences, was removed when analyses of LV GLS were performed by AI rather than by 2 different human readers using a semiautomatic reference method. These findings strongly support that the fast and reliable automated measurement of LV GLS provided by AI can improve echocardiographic assessment of LV function and should be considered for implementation in clinical practice.

The Clinical Implications of Improved Test-Retest Reproducibility in Repeated Echocardiograms

A reproducible and accurate evaluation of LV function is needed to provide optimal diagnosis and treatment to the individual patient. Correspondingly, changes or lack of changes in LV function are fundamental for clinical decision-making throughout the spectrum of heart diseases and constitute pillars for guideline-based decisions in patients with heart failure and valvular heart disease and in cardio-oncology.[16-18] Good within-patient test-retest reproducibility between repeated echocardiograms is therefore paramount for correct clinical decisions but is often overlooked in echocardiographic research. As the test and retest echocardiograms for each patient in our study were recorded without time delay at the same day, the differences between the 2 recordings relate to differences introduced by acquisitions or readings, and not real changes of LV function. Artificial intelligence may improve the ability to reveal true changes in LV function by removing the bias introduced by different readers. The many reader combinations of the present study resulted in a wide range of observed interobserver variability, which illustrates the importance of having multiple readers when reporting interobserver variability in clinical research.

Although the variability in assessment of LV function has been reported to be better with LV GLS compared with LVEF, reproducibility

Table 1 Study populations

Parameter	Data set I (n = 40)	Data set II (n = 32)
Demographics:		
Age, years	67 ± 11 (46-89)	60 ± 13 (28-88)
Gender, male	27 (68)	15 (47)
Body mass index, kg/m^2	27 ± 4 (19-34)	28 ± 4 (22-40)
Heart rate, bpm	74 ± 15 (44-132)	66 ± 10 (42 - 95)
Systolic blood pressure, mm Hg	142 ± 19 (100-176)	130 ± 17 (102-172)
Comorbidity, n (%)		
Hypertension	30 (75)	11 (34)
Coronary artery disease	29 (72.5)	0 (0)
Echocardiographic recordings:		
LVEF, %	52 ± 7 (34-65)	60 ± 5 (47-72)
End-diastolic volume, mL	100 ± 21 (58-156)	118 ± 26 (42-186)
End-systolic volume, mL	49 ± 14 (27-78)	47 ± 13 (18-88)
Frame rate, frames per second	73 ± 10 (50-95)	79 ± 6 (61-100)
Image quality recording 1:		
Good (0/18 segments missing)	20 (50)	13 (41)
Fair (1-2/18 segments missing)	12 (30)	10 (31)
Poor (>2/18 segments missing)	8 (20)	9 (28)
Image quality recording 2:		
Good (0/18 segments missing)	18 (45)	13 (41)
Fair (1-2/18 segments missing)	15 (37.5)	12 (37)
Poor (>2/18 segments missing)	7 (17.5)	7 (22)

Categorical data are presented as numbers n (%) and continuous data as mean ± SD (range).

might still be a major clinical challenge.[19,20] Ideally, serial measurements of any clinical metric should be performed by the same reader. However, in many clinical scenarios it is impractical or impossible to always have the same reader present for serial analyses. Compared with interobserver scenarios, the use of AI for repeated analyses of LV GLS reduced the MDC and mean absolute difference and removed the systematic bias, thus indicating improved reproducibility comparable to what could be achieved by repeated analyses by the same experienced reader.

Interpretation of Findings in the Context of Previous Studies

Even though there are commercially available fully automated methods for LV GLS measurements, we are not aware of any previous study evaluating test-retest reproducibility of such methods in repeated echocardiograms. It is expected that fully automated measurements in general have an advantage with respect to reproducibility, but whether the present findings could be extended to other fully automated methods must be evaluated in dedicated studies. Only a few studies have reported repeated echocardiogram test-retest performance of commercially available semiautomatic methods for LV GLS measurements, and with a wide range of variability.[20-24] Moreover, these studies were single center and measurements were performed by only 1 or 2 readers. Readers

Table 2 Test-retest reproducibility of GLS measurements for interreader, intrareader, and AI scenarios

Parameter	Data set I	Data set II
Mean absolute difference, strain units (%):		
Interreader scenarios, mean (range)	2.1* (1.5-3.4)	1.9* (1.4-2.6)
Intrareader scenarios, mean (range)	1.5 (1.4-1.6)	1.7 (1.4-2.0)
AI scenario	1.4	1.6
SE_M, strain units (%):		
Interreader scenarios, mean (range)	2.0* (1.3-2.8)	1.7* (1.2-2.4)
Intrareader scenarios, mean (range)	1.3 (1.2-1.5)	1.6* (1.2-2.0)
AI scenario	1.3	1.4
MDC, strain units (%):		
Interreader scenarios, mean (range)	5.5* (3.8-7.6)	5.2* (3.3-6.5)
Intrareader scenarios, mean (range)	3.7 (3.4-4.0)	4.5* (3.4-5.7)
AI scenario	3.7	3.9
Coefficient of variation, %:		
Interreader scenarios, mean (range)	9.6* (6.7-15)	9.2* (6.3-12.6)
Intrareader scenarios, mean (range)	7.0 (6.7-7.2)	8.7 (6.6-11.1)
AI scenario	7.2	8.5
ICCs:		
Interreader scenarios, mean (range)	0.72 (0.49-0.85)	0.63 (0.41-0.81)
Intrareader scenarios, mean (range)	0.84 (0.82-0.88)	0.67 (0.52-0.78)
AI scenario	0.84	0.70

*Significant difference (P < .05) between the mean and the AI scenario.

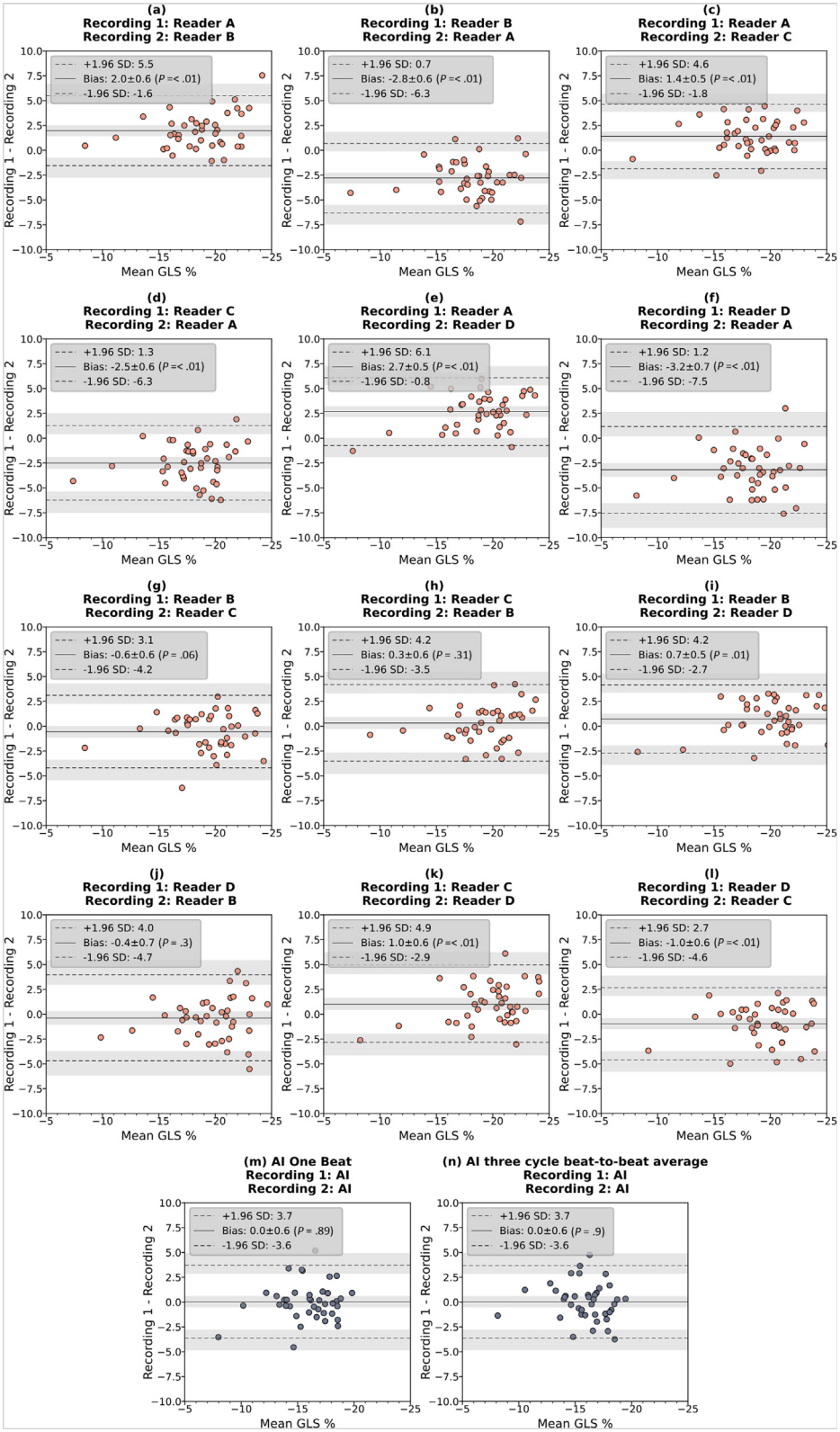

Figure 2 Data set I: test-retest interreader scenarios. Bland-Altman plots presenting bias and limits of agreement for the 12 interreader scenarios constructed by measurements made by readers A, B, C, and D using the semiautomatic reference method **(A–L)** and for the AI scenario without manual input **(M** and **N)**. The *gray shaded areas* represent the 95% CI of estimates.

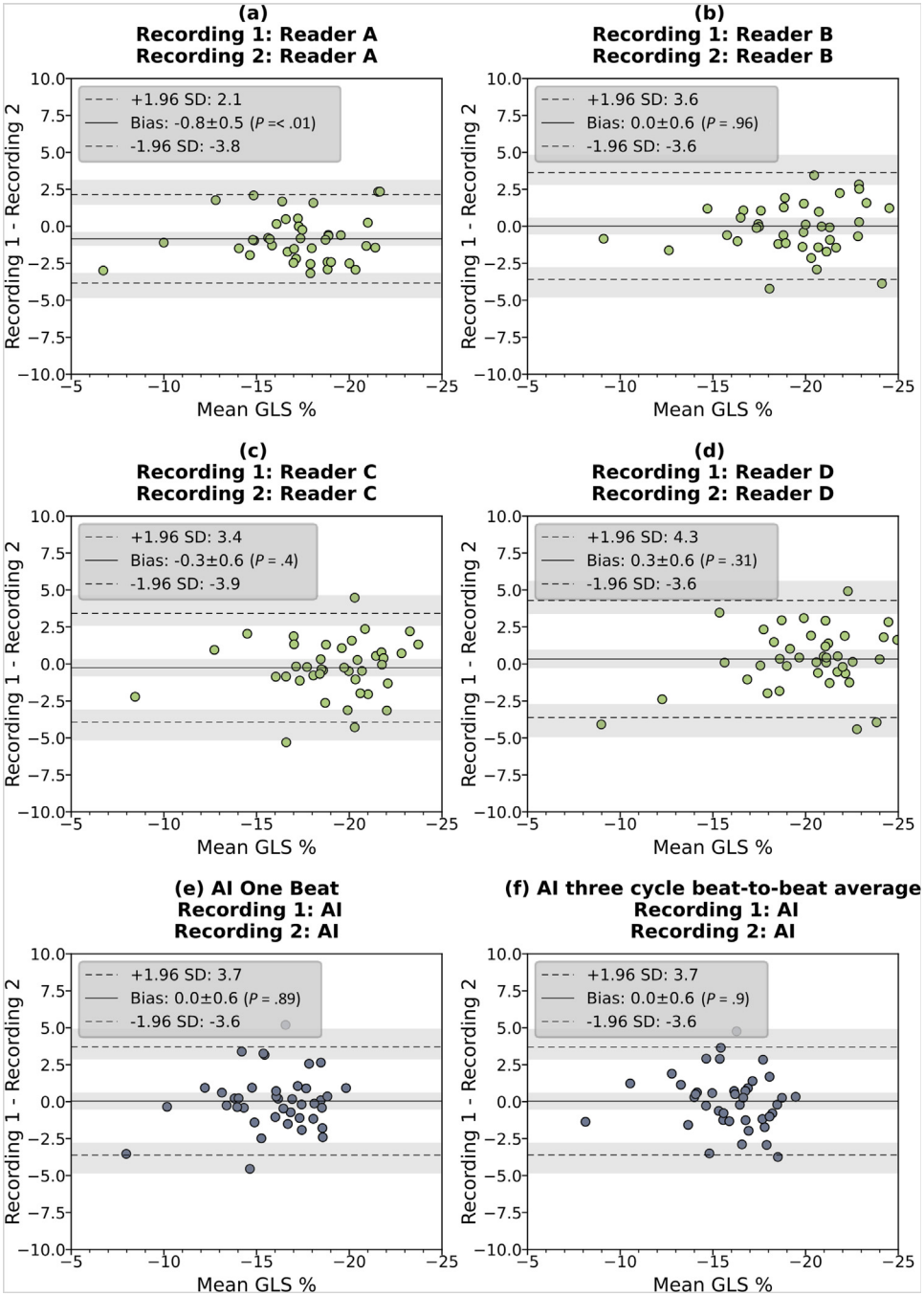

Figure 3 Data set I: test-retest intrareader scenarios. Bland-Altman plots presenting bias and limits of agreement for the 4 test-retest scenarios constructed when the same reader A, B, C and D analyzed both the first and second image recording **(A–D)** and for the AI test-retest scenario without manual input **(E** and **F)**. The *gray shaded areas* represent the 95% CI of estimates.

trained at different institutions may have slightly different conventions for how to perform manual adjustments when using the semiautomatic method for GLS measurements. Thus, the variability presented in our study may be more representative of the everyday clinic.

Variability in LV GLS between readers may have several contributing factors. By visual inspection of the ROIs extracted from the semiautomated method it seemed that the initiation of the ROIs was important for whether the endocardium and trabeculae as opposed to the myocardium were tracked. This was supported by quantification of the length of the ROI midline and the ventricular length, which differed significantly between observers. This tendency seemed to be particularly prominent in the apical region (Supplemental Online Figure 1) and more pronounced with less manual adjustment of the ROIs. Results for absolute strain values were on average higher for the reader who systematically positioned the ROI closer to the LV cavity. In a scenario where all readers were using the AI method, variations in LV GLS due to individual differences in ROI initiation would have been eliminated. These findings illustrate some of the benefits of standardization of measurements

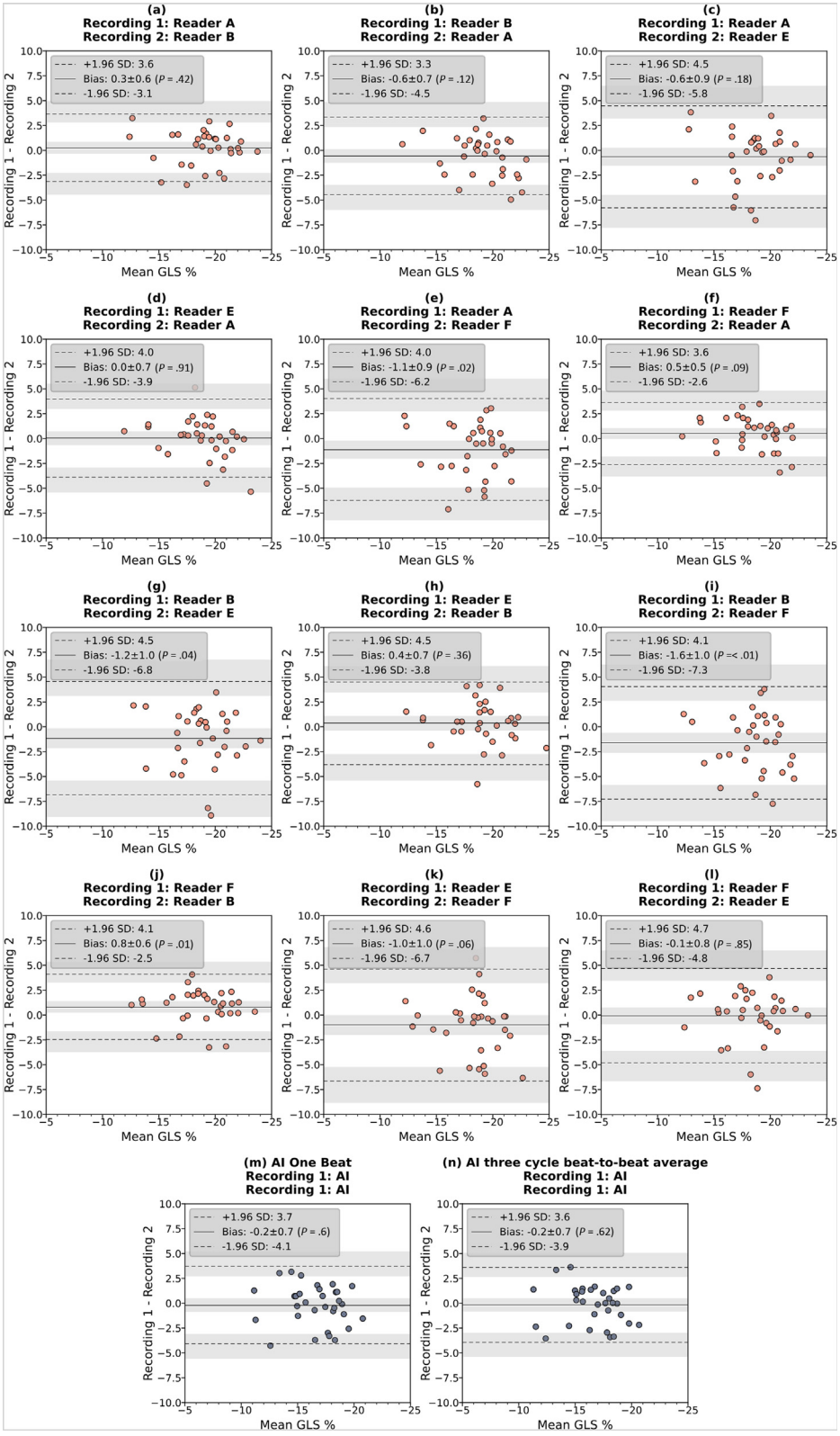

Figure 4 Data set II: test-retest interreader scenarios. Bland-Altman plots presenting bias and limits of agreement for the 12 interreader scenarios constructed by measurements made by readers A, B, E, and F using the semiautomatic reference method **(A–L)** and for the AI scenario without manual input **(M** and **N)**. The *gray shaded areas* represent the 95% CI of estimates.

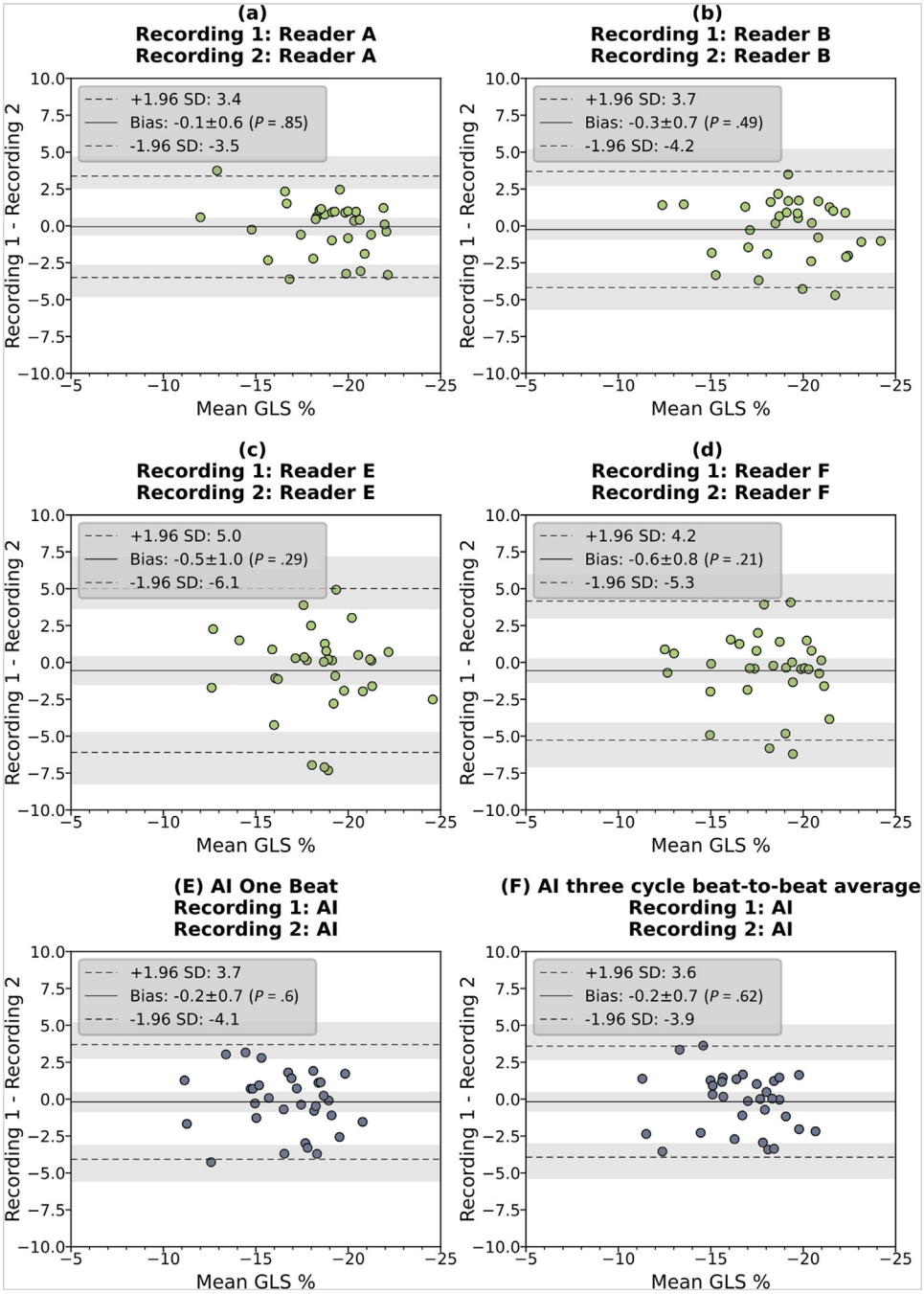

Figure 5 Data set II: test-retest intrareader scenarios. Bland-Altman plots presenting bias and limits of agreement for the 4 test-retest scenarios constructed when the same reader A, B, E, and F analyzed both the first and second image recording **(A–D)** and for the AI test-retest scenario without manual input **(E** and **F)**. The *gray shaded areas* represent the 95% CI of estimates.

by automated AI software, even when the semiautomatic algorithms seem to track adequately.

The mean LV GLS measured by AI was in the lower range of what has previously been reported with strain measured by speckle-tracking. However, the difference between AI and the semiautomatic speckle-tracking method is in line with the differences previously observed between ultrasound systems[22] and also with the validation studies of the novel AI method.[9] Intervendor, intersoftware, and intermodality variability is a known issue in strain imaging, and slightly different normal ranges have been reported for different vendors and analysis packages. The mean LV GLS in the present paper also corresponds to previously reported relative change in apical-to-basal ventricular length and strain measured by tissue Doppler.[25] Thus, small differences between the different methods are expected.

Beat-to-Beat Assessment

Assessment of beat-to-beat reproducibility revealed similar MDCs for both readers using the semiautomatic method, whereas MDC was

lower for the AI method. This implies that the ability to identify subtle changes in LV GLS is improved by using AI.

An important advantage of the AI method is that averaged beat-to-beat measurements of LV GLS are easily calculated within seconds, whereas this is very time-consuming using a semiautomatic method. This advantage of the AI method could be of great benefit when performing measurements in patients with irregular rhythms such as atrial fibrillation, where it is recommended to perform averaged measurements of at least 5 cardiac cycles.[26]

Limitations

The examinations of data set I were acquired using an older generation ultrasound system than that used for data set II (Vivid 7 vs Vivid E95). The older ultrasound system may have produced lower image quality than the newer system, and this could contribute to the difference in results between data sets. However, older generation ultrasound systems are still widely used worldwide, and including a data set acquired by these scanners therefore improves the generalizability of the results.

The participating readers were all experienced in echocardiography, but with variable practice in strain imaging. However, intra-reader variability for the less experienced readers was not statistically inferior compared with the 2 most experienced readers, indicating that test-retest variability is an issue even within experienced observers. The readers' experience could therefore not explain why AI had less variability than semiautomatic measurements. Moreover, the level of experience by the observers resembles many echo laboratories, which the authors believe adds to the clinical relevance of this study.

The proposed AI software is vendor independent and could potentially be used to analyze images from any other ultrasound machine. However, in this study the same vendor was used for all image acquisitions and reference measurements, and the results can therefore not be generalized to other ultrasound systems without further validation. Moreover, there is no gold standard for measuring LV GLS, and thus, it was not possible to conclude whether the reference method or the AI method produced the most accurate estimate of LV function. Therefore, the aim was not to compare the values of LV GLS obtained by the AI method with those obtained by the semiautomatic method but rather to investigate the test-retest and beat-to-beat variability of the AI method. Even though the novel AI-based LV GLS method has demonstrated good agreement with reference and this study shows the benefits of the method with respect to reproducibility, data on the clinical accuracy and prognostic impact should be documented before large-scale clinical implementation. There are commercially available automated LV GLS methods. We are, however, not aware of studies reporting on the test-retest performance of these methods, and comparisons to the current method must be addressed in future work.

CONCLUSION

The novel and fully automated AI method based on deep learning successfully provided consistent within-patient test-retest measurements of LV GLS in repeated echocardiograms recorded by different echocardiographers. The AI method removed bias and reduced test-retest variability compared with the case where different readers used conventional semiautomatic methods to measure LV GLS. The fast performance and high feasibility of the AI method may allow for real-time strain calculations performed during echocardiographic acquisitions in the future, thereby facilitating implementation of LV GLS and improving the workflow in clinical echocardiography.

DATA AVAILABILITY STATEMENT

The data set is available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

We thank the sonographers Even Olav Jakobsen and Kris Soler for their efforts in image acquisitions and measurements.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://doi.org/10.1016/j.echo.2023.02.017.

REFERENCES

1. Klaeboe LG, Edvardsen T. Echocardiographic assessment of left ventricular systolic function. J Echocardiogr 2019;17:10-6.
2. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2021;42:3599-726.
3. Priori SG, Blomstrom-Lundqvist C, Mazzanti A, et al. 2015 ESC guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: I task Force for the Management of patients with ventricular Arrhythmias and the Prevention of Sudden cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). Eur Heart J 2015;36:2793-867.
4. Plana JC, Galderisi M, Barac A, et al. Expert consensus for multimodality imaging evaluation of adult patients during and after cancer therapy: a report from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2014;27:911-39.
5. Porter TR, Shillcutt SK, Adams MS, et al. Guidelines for the use of echocardiography as a monitor for therapeutic intervention in adults: a report from the American Society of Echocardiography. J Am Soc Echocardiogr 2015;28:40-56.
6. Collier P, Phelan D, Klein A. A test in Context: myocardial strain measured by speckle-tracking echocardiography. J Am Coll Cardiol 2017;69:1043-56.
7. Seetharam K, Raina S, Sengupta PP. The Role of artificial intelligence in echocardiography. Curr Cardiol Rep 2020;22:99.
8. Asch FM, Descamps T, Sarwar R, et al. Human versus artificial intelligence–based echocardiographic analysis as a predictor of outcomes: an analysis from the world Alliance Societies of echocardiography COVID study. J Am Soc Echocardiogr 2022;22:894-7317.
9. Salte IM, Ostvik A, Smistad E, et al. Artificial intelligence for automatic measurement of left ventricular strain in echocardiography. JACC Cardiovasc Imaging 2021;10:1918-28.
10. Østvik A, Salte IM, Smistad E, et al. Myocardial function imaging in echocardiography using deep learning. IEEE Trans Med Imaging 2021;5:1340-51.

11. Galderisi M, Cosyns B, Edvardsen T, et al. Standardization of adult transthoracic echocardiography reporting in agreement with recent chamber quantification, diastolic function, and heart valve disease recommendations: an expert consensus document of the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging 2017;18:1301-10.
12. Voigt JU, Pedrizzetti G, Lysyansky P, et al. Definitions for a common standard for 2D speckle tracking echocardiography: consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging. Eur Heart J Cardiovasc Imaging 2015;16:1-11.
13. Amundsen BH, Helle-Valle T, Edvardsen T, et al. Noninvasive myocardial strain measurement by speckle tracking echocardiography: validation against sonomicrometry and tagged magnetic resonance imaging. J Am Coll Cardiol 2006;47:789-93.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
15. Shieh G. The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. BMC Med Res Methodol 2018;18:45.
16. Maddox TM, Januzzi JL, Allen LA, et al. 2021 update to the 2017 ACC expert consensus decision Pathway for optimization of heart failure treatment: Answers to 10 Pivotal issues about heart failure with reduced ejection fraction. J Am Coll Cardiol 2021;77:772-810.
17. Otto CM, Nishimura RA, Bonow RO, et al. 2020 ACC/AHA guideline for the management of patients with valvular heart disease. J Am Coll Cardiol 2021;77:e25-197.
18. Thavendiranathan P, Negishi T, Somerset E, et al. Strain-guided management of potentially cardiotoxic cancer therapy. J Am Coll Cardiol 2021;77:392-401.
19. Chan J, Shiino K, Obonyo NG, et al. Left ventricular global strain analysis by two-dimensional speckle-tracking echocardiography: I learning Curve. J Am Soc Echocardiogr 2017;30:1081-90.
20. Baron T, Berglund L, Hedin EM, et al. Test-retest reliability of new and conventional echocardiographic parameters of left ventricular systolic function. Clin Res Cardiol 2019;108:355-65.
21. Costa SP, Beaver TA, Rollor JL, et al. Quantification of the variability associated with repeat measurements of left ventricular two-dimensional global longitudinal strain in a real-world setting. J Am Soc Echocardiogr 2014;27:50-4.
22. Farsalinos KE, Daraban AM, Unlu S, et al. Head-to-Head comparison of global longitudinal strain measurements among nine different vendors: I EACVI/ASE inter-vendor comparison study. J Am Soc Echocardiogr 2015;28:1171-1181,e2.
23. Thorstensen A, Dalen H, Amundsen BH, et al. Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study. Eur J Echocardiogr 2009;11:149-56.
24. Karlsen S, Dahlslett T, Grenne B, et al. Global longitudinal strain is a more reproducible measure of left ventricular function than ejection fraction regardless of echocardiographic training. Cardiovasc Ultrasound 2019;17:18.
25. Dalen H, Thorstensen A, Aase SA, et al. Segmental and global longitudinal strain and strain rate based on echocardiography of 1266 healthy individuals: the HUNT study in Norway. Eur J Echocardiogr 2010;11:176-83.
26. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015;28:1-39.e14.