DANSK TIDSSKRIFT FOR AKUTMEDICIN

DANISH JOURNAL OF EMERGENCY MEDICINE

Dansk Tidsskrift for Akutmedicin

Vol. 8, No. 1

Accepted: 17 July 2025 | Published: 18 July 2025

DOI: 10.7146/akut.v8i1.149710

ORIGINAL ARTICLE

OPEN ACCESS

Automatic Ejection Fraction Agreement Between Handheld and Midrange Ultrasound Devices

Meryem Hamodi[1] | Annmarie Touborg Lassen [2] | Stefan Posth [2*]

[1]University of Southern Denmark, Odense, Denmark [2]Department of Emergency Medicine, Odense University Hospital, Odense, Denmark

Correspondence (*): stefan.posth@rsyd.dk

Abstract

Background

The integration of artificial intelligence (AI) in key cardiac function parameters, such as left ventricular ejection fraction (LVEF), can hold important value for clinicians, both in terms of time consumption and interobserver variability. However, the reproducibility between devices remains unknown.

Aim

The purpose of this study was to assess two ultrasound devices with their automated LVEF (auto-LVEF) measurements: the midrange GE venue (GEv), and the handheld Butterfly iQ+(Bfi); regarding correlation in ejection fraction (EF), time consumption, and image quality (IQ).

Method

Adult emergency room patients were included and scanned using both ultrasound devices by a novice operator. In each case, the objective was to acquire an apical four-chamber view and calculate the EF with each device's pre-installed AI software. Out of those, 12 patients were rescanned by a physician experienced in cardiac ultrasound to evaluate the interoperator agreement.

Results

A total of 150 patients were included, with a median age of 64 years; 51% were female. The GEv and Bfi successfully generated auto-EF measurements in 73% (95% confidence interval [CI]: 65%–80%) and 52% (95% CI: 44–60%) of cases, respectively. The agreement in EF measurements between the GEv's real-time EF and the Bfi's Simpson monoplane method was high with a correlation coefficient r = 0.70 (0.60–0.77), $p < 0.001$. Bland-Altman analysis demonstrated a bias of 0.84% (95% upper and lower limits of agreement: 15.0% and -13.3%). The median scanning time in both apparatuses was 2 minutes (IQR GEv 1–2, IQR Bfi 1–3), the median IQ score was 4/5 (IQR 4–5) in GEv and 3.5/5 (IQR 3–4) in Bfi. The interobserver agreement was high, with a Kappa of $\kappa_{GEv} = 0.75$ and $\kappa_{Bfi} = 0.82$.

Conclusion

In conclusion, Bfi had a lower success rate in calculating EF and a lower IQ than GEv. However, when auto-EF was successfully obtained, a strong correlation was observed between the machines.

Keywords: Emergency medicine; Cardiac ejection fraction; Emergency department; Point-of-care ultrasound; Artificial intelligence

Introduction

The left ventricular ejection fraction (LVEF) is one of the most important indices for assessing cardiac function. Echocardiography and focused cardiac ultrasound are the main imaging modalities for measuring ejection fraction (EF), due to their low cost, non-invasiveness, and accessibility, making them an essential diagnostic tool in emergency settings [1]. However, manual EF assessment is both time-consuming and subject to operator variability [2].

Machine learning and artificial intelligence (AI)-enabled automated algorithms are emerging as promising solutions for these challenges. Most importantly, in helping to ensure the quality and consistency in visual/semi-quantitative measurements, minimise misreads, as well as train the clinicians' eyeballing ability. As a result, AI-based algorithms are gaining a larger audience among medical professionals seeking to exploit their usability and accuracy [3]. Some AI algorithms account for both regional and global wall motion abnormalities when estimating EF, but far from all offer direct integration.

A variety of ultrasound machines incorporating AI algorithms are now available on the market from various manufacturers, including handheld devices. These portable devices compete with traditional cart-based machines, since they are smaller and cheaper, but come with certain limitations. This study aims to evaluate the agreement between a midrange device and a handheld device in terms of correlation in EF measurements, image quality (IQ), and time efficiency. In this setup, we did not compare the automated EF measurement to a gold standard.

HVAD VED VI?

Brugen af point-of-care ultralyd i akutmodtagelser bliver mere og mere udbredt, og mange ultralydsapparater indeholder allerede nu AI-moduler.

Fact box (in Danish)

Method

STUDY DESIGN

Adult patients admitted to the emergency room at Odense University Hospital were invited to participate in the study, regardless of their presenting symptoms. The study was conducted from March to May 2023. All participants provided both verbal and written informed consent before undergoing two ultrasound scans: one with a midrange device and one with a handheld device. The only exclusion criterion was the inability to provide fully informed consent. Patients were only scanned when the examiner was on shift in the emergency room.

STUDY APPROVAL

The study protocol was submitted to the Ethics Committee of the Region of Southern Denmark, which determined that formal approval was not required. Written informed consent was obtained from all participants. Permission for data storage was granted by the Region of Southern Denmark (J. No. 22/59027).

ULTRASOUND

The ultrasound devices being assessed in this study were the conventional, midrange ultrasound device “GE venue” (GEv), and the smaller handheld device “Butterfly iQ+” (Bfi, released in 2020). The GEv has several probes which, like those in conventional ultrasound devices, contain piezoelectric crystals. Recommended for cardiac scanning is the phased array transducer. The Bfi has a single probe that emulates any type of transducer and, instead of using piezoelectric crystals, contains a silicon chip that converts voltage into resonance (Figure 1). The algorithms used by the two devices to assess LVEF are based on similar parameters. The GEv uses Real-Time EF (GE HealthCare, Chicago, IL, USA), which is an auto-EF calculation tool based on the apical 4-chamber view (A4CV) that traces the walls of the left ventricle to identify end-systolic and end-diastolic frames.

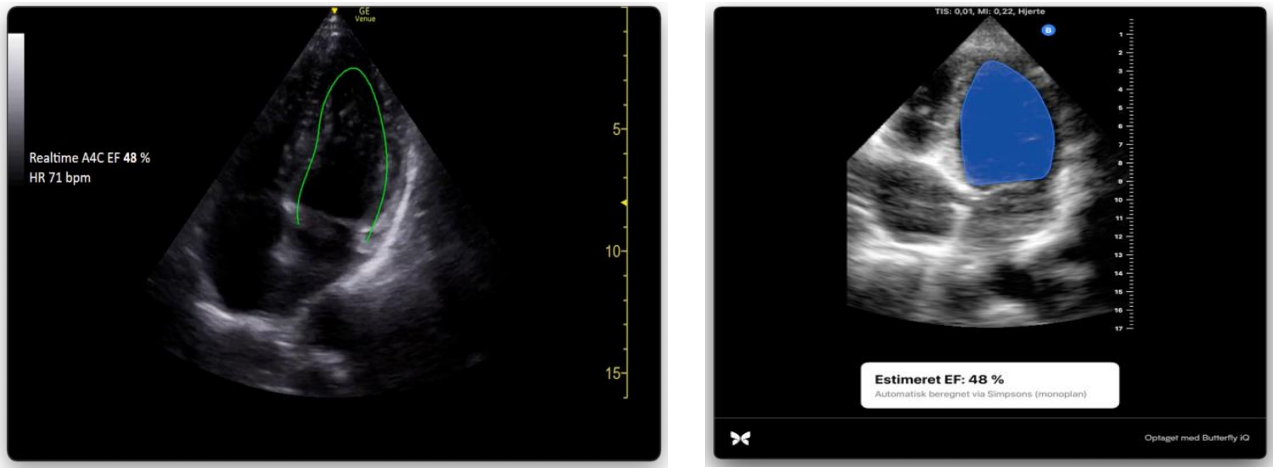

Figure 1. GE venue's Real-Time EF (left) and Butterfly iQ's Simpson's EF screen layouts (right)

This real-time EF is an AI-enabled tool that continuously calculates the EF during live scanning in A4CV to provide an immediate result, granted that the quality of the images is sufficient, which is determined by a colour-based quality indicator. There are three levels on the quality indicator scale: green, yellow, and red, corresponding to high, medium, or low view quality, based on scanning quality, the tool's ability to identify the A4CV, and the EF result's consistency [4]. The Bfi is connected to a smartphone and uses a software application (Butterfly Network, Inc., Burlington, MA, USA) called Simpson's EF. It uses Simpson's monoplane method, an automatic EF tool that allows for calculating the LVEF, also based on the A4CV. A key difference is that the calculation is based on a 3-second-long segment instead of real-time, and after the segment is analysed, an auto-EF is given. Similarly to GEv, it uses a colour scale to indicate the IQ, from green to red, where green indicates high and red indicates low quality [5].

The scans were performed by an ultrasound novice, a sixth-year medical student. The student was trained online and by an ultrasound expert. The student was certified in focus-assessed transthoracic echocardiography (FATE) after 35 on-site scans. In addition, a POCUS (point-of-care ultrasound) specialist assessed a fraction of the study total: 12 patients (approximately 10% of the total), for interobserver agreement. The POCUS specialist was certified in FATE and had performed more than 1,000 FATE examinations. The scans were performed so that the operator shifted between GEv and Bfi with each scan, making sure to start with each machine every other time, to ensure a fair time in image acquisition, since less time is consumed to locate the heart the second time around.

DATA COLLECTION

The data that was collected from the patients were: patients' personal registration number, sex, age, weight, height, BMI, participants' preliminary incoming diagnosis, and LVEF. All information was stored in REDCap [6].

Image quality was assessed based on how well-defined the heart chambers were and the overall acuity of the image. An image quality score from 1 to 5 was given, where 1 was very poor quality, making the heart unidentifiable (black screen); 2 was poor quality (heart contours just visible); 3 was acceptable quality; 4 was good quality; and 5 was perfect quality (equivalent to a textbook presentation). The time per scanning was recorded on a phone; starting from the placement of the probe on the patient's chest, ongoing while the clinician identified the optimal A4CV, and ending when the auto-EF measurement is obtained. Time was recorded in full minutes. The auto-EF measurements made with GEv were occasionally oscillating in character; in such cases, an average value of the EF was noted.

STATISTICAL ANALYSIS

Data were presented as a median and interquartile range (IQR) for variables with a non-normal distribution, while the mean and standard deviation (SD) were used, when appropriate, for normally distributed data. Categorical variables are expressed as frequencies and percentages/proportions. The correlation in auto-EF between GEv and Bfi was described using a scatter plot with linear regression, and the correlation coefficient was calculated. A Bland-Altman plot was generated to visualise the agreement and identify systemic errors (bias). The interobserver agreement was evaluated using Cohen's kappa coefficient, interpreted according to the following criteria: <0.00 (no/poor agreement), 0.01–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (near-perfect to perfect agreement) [7]. A p-value ≤ 0.05 was considered statistically significant.

Results

We included 150 participants (76 females, median age: 64, IQR 48–76) from March to May 2023. Of these, 108 participants were admitted to the hospital with a preliminary medical incoming diagnosis, such as dehydration, pneumonia, or COPD in exacerbation, while 42 were surgical patients with conditions such as gallstones, appendicitis, ileus, etc. (Table 1). The diagnoses are presented in a diagram (Supplement 1).

HVAD UNDERSØGER STUDIET?

Forskellen mellem billedkvalitet, tidsforbrug og kvaliteten af den AI-beregnede EF mellem to hyppigt brugte apparater, udført af en novice og en erfaren operatør

Fact box (in Danish)

Auto-EF measurement was successful in 110 cases (73%, 95% CI 65%-80%) using GEv and in 78 cases (52%, 95% CI 44%-60%) using Bfi (Table 2).

The median EF measurements in GEv and Bfi were 55% (IQR 49–59) and 56% (IQR 48–60), respectively. The correlation coefficient between GEv and Bfi was strong, r = 0.70 (95% CI: 0.60–0.77, p < 0.001), with a regression equation: $f(x) = 0.97x + 1.05$. Bland-Altman analysis showed a mean difference (bias) of 0.84 % (95% limits of agreement were -13.3% and 15.0%), indicating a slight overestimation in GEv measurements (Fig. 2). No proportional or systemic bias was identified.

The median time per scan in GEv and Bfi was 2 minutes for both devices (IQR: GEv 1–2, Bfi 1–3). Image quality, assessed by the operator on a 1–5 scale, was rated as good for GEv (median 4, IQR 4–5) and as okay-to-good for Bfi (median 3.5, IQR 3–4) (Fig. 3).

Interobserver agreement between Operators 1 novice) and 2 (expert) - (Op1 and Op2) was substantial, Cohen's kappa values of κ_{GEv} = 0.75 and κ_{Bfi} = 0.82 (Supplement 2).

Table 1. Patients' characteristics with GEv and Bif. *n*, number of patients; BMI, body mass index; IQR, interquartile range; IQ, image quality.

Patients' characteristics	Total	EF GEv successful	EF Bfi successful	EF GEv unsuccessful	EF Bfi unsuccessful
Median age, years (IQR)	64 (48–76)	61.5 (43–74)	60.5 (32–72)	67.5 (61–76)	68 (56–80)
Max./Min. age, years	20/94	20/92	20/92	24/94	24/94
Sex distribution, female, %	51	49	49	55	53
Median BMI (IQR)	26 (23–32)	24.7 (23–29)	24 (22–27)	31.4 (25–36)	30.5 (25–35)
Min./Max. BMI	14.9/53.9	14.9/51.3	14.9/40.8	18.1/53.9	18.1/53.9
Heart failure, *n*	2	2	1	0	1
Atrial fibrillation, *n*	2	2	1	0	1
Breast implant, *n*	3	1	1	2	2
Deviated heart axis	8	3	1	5	7

Surgical/Medical patients, %	72/28	70/30	71/29	77.5/22.5	74/26
Patient positioning, left lateral decubitus, 45-90º	132	101	73	31	59
Patient positioning, supine/<30º	18	9	5	9	13

Table 1 (continued).

Scannings' characteristics	
EF possible GEvR, %	73 (95% CI 65–80)
EF possible Bfi, %	52 (95% CI 44–60)
Median EF measurement GEvR (IQR)	55 (49–59)
Median EF measurement Bfi (IQR)	56 (48–60)
Median time GEvR, minutes (IQR)	2 (1–2)
Median time Bfi, minutes (IQR)	2 (1–3)
Median picture quality GEvR/Bfi successful	5/4
Median picture quality GEvR/Bfi unsuccessful	3/3
Correlation coefficient GEv and Bfi, r	0.7 (95% CI: 0.60–0.77, $p < 0.001$)

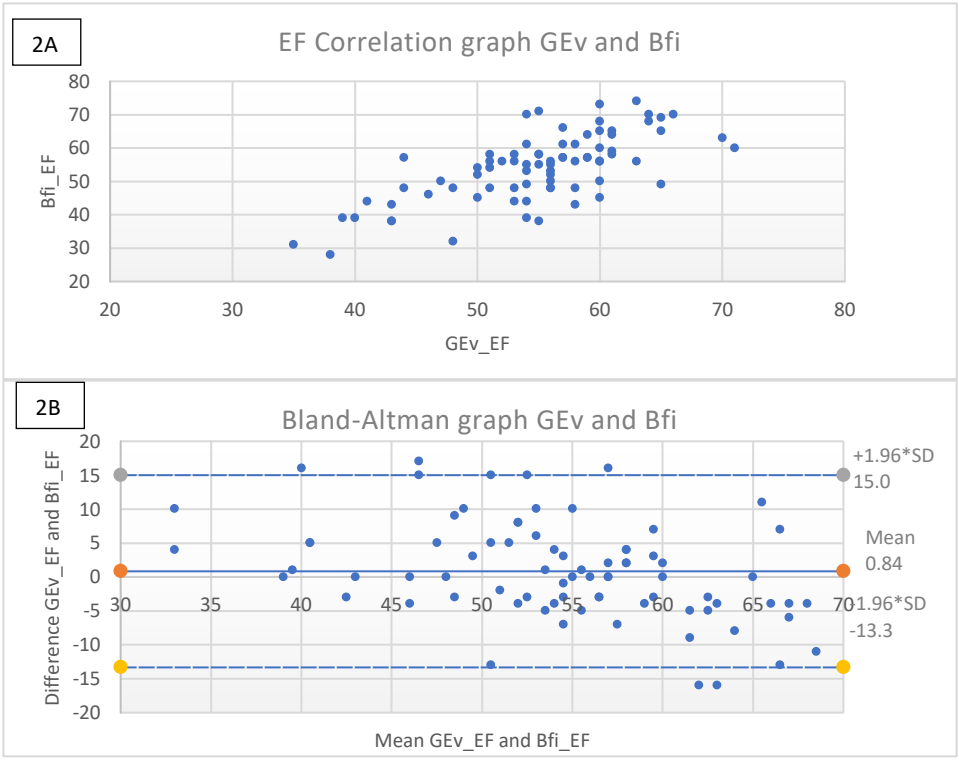

Figure 2. Agreement between machine learning-based automated ejection fraction measurements using GE Venue (GEv) and Butterfly iQ (Bfi): correlation graph (A) and Bland-Altman analysis (B).

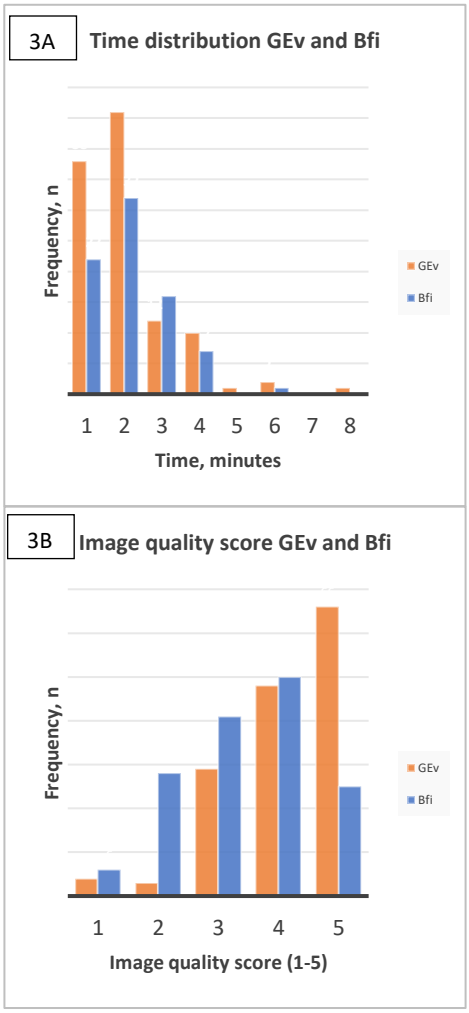

Figure 3. Time distribution (A) and image quality score graphs comparing GE Venue and Butterfly iQ.

Discussion

SUMMARY OF EVIDENCE

To our knowledge, this is the first study to evaluate the agreement in auto-LVEF between handheld and mid-range ultrasound devices. Our findings indicate a strong correlation between GEv and Bfi in auto-LVEF measurements. Additionally, GEv demonstrated superior image quality compared to Bfi, while the time required to identify and calculate EF was similar for both devices.

COMPARISON OF STUDIES AND DEVICES

There is growing evidence supporting the feasibility of machine learning programs in POCUS and echocardiography [8]. A few studies have experimentally developed and evaluated neural network models for fully automated EF, yielding promising results [9-12]. However, clinical validation studies remain scarce, particularly for mid-range ultrasound devices, while the so-called pocket devices are still an unexplored area. One study assessed GEv's real-time auto-EF in critically ill patients and found a strong correlation between the auto-EF and conventional measurement [13]. Compared to high-end ultrasound devices, handheld ones generally have a lower IQ and smaller screen sizes but offer advantages such as portability and affordability. AI-powered algorithms are expected to enhance time efficiency and operator-independent calculations, helping inexperienced users in LVEF assessment. With this in mind, our agreement study was conducted by an ultrasound novice.

The success of the automated EF algorithms in both GEv and Bfi appeared to be strongly correlated with image quality. In our study, Bfi successfully generated auto-EF in 78 patients (52%), similar to findings by *Bacariza et al., 2023*, where less than half of the cases had successful auto-EF results [14]. Participants with just acceptable or suboptimal IQ (≤3/5) had unsuccessful auto-EF measurements in 70% and 86% of cases in GEv and Bfi, respectively. Two patients with excellent IQ in GEv still had unsuccessful auto-EF measurements due to a deviated heart axis, making it difficult to obtain an optimal apical 4-chamber view (A4CV). This suggests that distorted cardiac positioning may hinder the feasibility of automated EF assessment. Both GEv and Bfi use single-plane imaging, which is known to have lower precision and a tendency to underestimate EF compared to biplane methods [15]. Overall, we observed lower EF estimates with Bfi than with GEv. Since the auto-EF measurement made with GEv is based on real-time calculations, measurements occasionally fluctuated, and a mean value was recorded in such cases. This variability may pose a challenge to the untrained eye of a novice, while an expert might "eyeball" the EF automatically. In contrast, Bfi analyses a single captured segment, potentially obscuring such dynamic variations. However, in cases with poor IQ or suboptimal A4CV acquisition, Bfi occasionally produced erroneous auto-EF values, such as negative EF values (e.g., -9%).

Regarding patient positioning, while the optimal position is lying on the left side to improve heart visibility, many emergency patients are unable to assume this posture, resulting in suboptimal images and a higher failure rate for EF calculation. Additional challenges for auto-EF measurements included obesity and large breast tissue. Patients with breast implants also posed difficulties for the algorithms. Conversely, lean patients and those who had undergone mastectomies were easier to image, resulting in a higher success rate for auto-EF calculations.

We found a strong correlation between auto-EF measurements from GEv and Bfi, at the individual patient level. In comparison, the average interuser variation among cardiologists is approximately 10% [16]. Regarding the interobserver agreement, we found only minor differences between the novice and the POCUS specialist.

STRENGTHS AND LIMITATIONS

This study directly compares measurements from the two devices rather than evaluating them against a gold standard. Our findings demonstrate the correlation of these measurements in emergency department patients.

We consider it a strength of our study that the patients consisted of a heterogeneous group with an almost equal male-to-female ratio. However, certain limitations must be acknowledged. Although the population was diverse in some aspects, the range of EF measurements was relatively narrow and within a normal spectrum, which may limit the generalizability of our findings to patients with severely reduced EF.

Since not all included patients had a clinical indication for scanning, it remains unknown whether the agreement between methods would hold in a population with more cases at the lower end of the EF spectrum. Furthermore, this is a small, single-centre study with a limited number of participants, most of whom were clinically stable and mostly able to cooperate. No official guidelines were used for IQ assessment; the scale was created and defined by the operators. Additionally, all scans were performed by the same novice operator, with only a small subset of participants rescanned by the experienced ultrasound operator, resulting in low statistical power for interobserver agreement.

Furthermore, GE's real-time AI algorithm for EF estimation does not incorporate the wall motion score index (WMSI) in its calculation. WMSI is a separate method used to c regional wall motion of the left ventricle, which aids in evaluating the extent of myocardial dysfunction.

PERSPECTIVES

Despite the auto-EF limitations identified in our study, these results could change with rapid future advances in AI and may adumbrate promising value for clinical usability. Further research could help galvanise manufacturers and developers to refine and upgrade both the devices and their algorithms, ultimately improving accuracy and reliability in clinical settings.

HVAD TILFØJER STUDIET?

AI kan ikke måle EF i alle tilfælde, og billedkvaliteten fra det håndholdte apparat er nogle gange utilstrækkelig.

Fact box (in Danish)

Conclusion

This study demonstrated that Bfi had a lower success rate in calculating an EF value and a lower IQ than GEv. However, when auto-EF was successfully obtained, a strong correlation between the two devices was observed.

Abbreviations

LVEF – left ventricular ejection fraction; EF – ejection fraction; GEv – GE venue; Bfi – Butterfly iQ; IQ – image quality; A4CV – apical 4-chamber view.

Funding

No specific grant was provided for this research. The ultrasound apparatuses used in our study belong to the hospital's emergency department.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics approval and consent to participate

The study protocol was presented to the Ethics Committee of the Region of Southern Denmark, which determined that formal approval was not required. Written informed consent was obtained from each patient. Permission to store data was given by the Region of Southern Denmark (J. No. 22/59027).

References

1. 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. Eur Heart J Cardiovasc Imaging. 2015;16(3):233-70. doi: 10.1093/ehjci/jev014.

2. Spahillari A, Colbert JA, Fox J, Singh J, Nayyar P, Chen S, et al. On-call transthoracic echocardiographic interpretation by first year cardiology fellows: comparison with attending cardiologists. BMC Med Educ. 2019;19(1):213. doi: 10.1186/s12909-019-1645-5.

3. 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(9):e009303. doi: 10.1161/CIRCIMAGING.119.009303.

4. Hila Best PU. GE Healthcare. Venue™ Family Real-Time EF. 2022. Available from: https://www.gehealthcare.dk/. doi: not available. [Access date 19. september 2024].

5. Butterfly Network. Calculate Simpson's Ejection Fraction. 2022. Available from: https://www.butterflynetwork.com/. doi: not available. [Access date 19. september 2024].

6. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377-81. doi: 10.1016/j.jbi.2008.08.010.

7. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360-3. doi: not available.

8. Barry T, Odreman L, Jain S, Campbell M, Nakatani S, Mihailescu SD. The role of artificial intelligence in echocardiography. J Imaging. 2023;9(2):40. doi: 10.3390/jimaging9020040.

9. Jafari MH, Habijan M, LeBel A, Tsai K, Fenster A. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. Int J Comput Assist Radiol Surg. 2019;14(6):1027-37. doi: 10.1007/s11548-019-01955-9

10. Dadon Z, Eitan A, Raz A, Freund Y, Levy G. Use of artificial intelligence as a didactic tool to improve ejection fraction assessment in the emergency department: a randomized controlled pilot study. AEM Educ Train. 2022;6(2):e10738. doi: 10.1002/aet2.10738.

11. Moal O, Brunet M, Couronne R, D'Hooge J, Bernard O. Explicit and automatic ejection fraction assessment on 2D cardiac ultrasound with a deep learning-based approach. Comput Biol Med. 2022;146:105637. doi: 10.1016/j.compbiomed.2022.105637.

12. Asch FM, Jankowski M, Cleve J, Adams M, Markuzon N, Manovel A, et al. Deep learning-based automated echocardiographic quantification of left ventricular ejection fraction: A point-of-care solution. Circ Cardiovasc Imaging. 2021;14(6):e012293. doi: 10.1161/CIRCIMAGING.120.012293.

13. Varudo R, Bento L, Vitorino R, Vieira J, Miro V, Guerreiro R, et al. Machine learning for the real-time assessment of left ventricular ejection fraction in critically ill patients: a bedside evaluation by novices and experts in echocardiography. Crit Care. 2022;26(1):386. doi: 10.1186/s13054-022-04255-4.

14. Bacariza J, Sousa C, Teixeira C, Pinto J, Ribeiro L, Martins R, et al. Smartphone-based automatic assessment of left ventricular ejection fraction with a silicon chip ultrasound probe: a prospective comparison study in critically ill patients. Br J Anaesth. 2023;130(6):e485-7. doi: 10.1016/j.bja.2023.02.014.

15. Heinen A, Schulte J, Scherschel K, Meyer C, Kelm M, Kleinbongard P, et al. Echocardiographic analysis of cardiac function after infarction in mice: validation of single-plane long-axis view measurements and the biplane Simpson method. Ultrasound Med Biol. 2018;44(7):1544-55. doi: 10.1016/j.ultrasmedbio.2018.03.008.

16. Johri AM, Picard MH, Newell J, Marshall JE, King MEE, Hung J. Can a teaching intervention reduce interobserver variability in LVEF assessment: a quality control exercise in the echocardiography lab. JACC Cardiovascular imaging [Internet]. 2011 Aug 1 [cited 2023 Apr 17];4(8):821–9. doi: 10.1016/j.jcmg.2011.06.004.