openheart

User experience and image quality influence on performance of automated real-time quantification of left ventricular function by handheld ultrasound devices: a diagnostic accuracy study with data from general practitioners, nurses and cardiologists

Anna Katarina Hjorth-Hansen [ID] , Malgorzata Izabela Magelssen, Garrett Newton Andersen, Torbjørn Graven, Jens Olaf Kleinau, Bodil Landstad, Lasse Løvstakken, Kyrre Skjetne, Ole Christian Mjølstad, Havard Dalen [ID]

Correspondence to
Dr Anna Katarina Hjorth-Hansen; annakh@ntnu.no

ABSTRACT

Background and objectives Echocardiography is the cornerstone of heart failure (HF) diagnosis, but expertise is limited. Non-experts using handheld ultrasound devices (HUDs) challenge the clinical yield. Left ventricular (LV) ejection fraction (EF) is used for assessment and grading of HF. Mitral annular plane systolic excursion (MAPSE) reflects LV long-axis shortening. Automatic tools for quantification of EF (autoEF) and MAPSE (autoMAPSE) are available on HUDs. We aimed to explore the importance of user experience and image quality for autoEF and autoMAPSE on HUDs, and how image quality influences the feasibility, agreement and reliability in patients with suspected HF.

Methods General practitioners, registered cardiac nurses and cardiologists represented the novice, intermediate and expert users, respectively, in this diagnostic accuracy study. 2543 images were evaluated by an external, blinded cardiologist by a five-parameter, prespecified score (four-chamber view, LV alignment, apical mispositioning, mitral annular assessment and number of visible endocardial segments) graded 0–6.

Results Feasibility was higher with increasing image quality. In all recordings, irrespective of user, the average image quality score and the five prespecified scores were associated with the feasibility of autoEF and autoMAPSE (all p<0.001). Image quality was more important for the feasibility of autoMAPSE than autoEF. Image quality was not important for the agreement of autoEF (R^2 2%) and autoMAPSE (R^2 7%). Combining all user groups, the reliability was lower with larger within-patient variability in image quality of the repeated recordings (p≤0.005). Similar associations were not found in user group specific analyses (p≥0.16). Patients' characteristics were only weakly associated with image quality score ($R^2 ≤ 4\%$).

Discussion Image quality was important for feasibility but does not explain the low agreement with reference or the modest within-patient reliability of automatic decision-support software on HUDs for all user groups in patients with suspected HF.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Handheld ultrasound devices (HUDs) have been used by non-experts for a decade, while decision support software to aid in evaluation of cardiac function is recently introduced.

⇒ In patients with suspected heart failure, we aimed to study how user experience and image quality influenced automatic quantification of left ventricular function with respect to feasibility, agreement and reliability.

WHAT THIS STUDY ADDS

⇒ Image quality was positively associated with feasibility across inexperienced general practitioners, intermediate experienced registered cardiac nurses and experienced cardiologists. However, image quality did not explain a modest agreement and reliability of the automatic decision support software for quantification of ejection fraction and mitral annular excursion.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Further refinement of the automatic decisionsupport software is needed before implementation into clinical practice.

INTRODUCTION

Echocardiography is the cornerstone for diagnosis and follow-up of heart failure (HF), but echocardiographic expertise is limited

to a few selected occupational groups. Left ventricular (LV) ejection fraction (EF) is widely used for assessing HF and grading of the severity.[1] Furthermore, mitral annular plane systolic excursion (MAPSE) is a sensitive and robust measure reflecting LV long-axis shortening and is less dependent on echogenicity.[2] Image quality is a mandatory prerequisite for a correct diagnosis by echocardiography where patient specific factors (eg, obesity, hyperinflated lungs and arrhythmias) will impair the acoustic environment and complicate image acquisition irrespective of the user's experience.[2] In addition, the users' experience may interfere with image quality.[3] Non-experts commonly use hand-held ultrasound devices (HUDs), which challenges the clinical yield of ultrasound diagnostics.[4] Advances in user support as real-time automatic measurements of cardiac structure and function may improve the diagnostic yield for non-experts, but initial results are conflicting.[5–7] How user characteristics and the quality of the recorded images influence the feasibility and reliability of automatic decision support software is not well known. To our knowledge, no study has explored the importance of user experience and image quality for quantification of EF (autoEF) and mitral annular systolic plane excursion (autoMAPSE) by HUDs.

The aim of this study was to evaluate how the users' experience may influence image quality of HUD recordings.

In addition to how various categories of image quality may influence the feasibility, agreement and reliability of real-time automatic decision support software for quantification of LV function by HUDs.

METHOD
Population and study design
Patients with suspected HF referred for cardiac evaluation at Levanger Hospital, Norway, were invited to take part in the study. Exclusion criteria were age <18 years, previous cardiac imaging within the last decade and known HF. The inclusion period was between June 2018 and May 2020, including a pause from March to May 2020 due to the COVID-19 pandemic. All participants gave their informed, verbal consent prior to inclusion and written consent the day of inclusion.

The general practitioners (GPs) were selected by the municipality administration based on their previously defined positions in municipality organised healthcare. The registered cardiac nurses (RCNs) were chosen according to their position at the outpatient HF clinic at Levanger Hospital, where all available RCNs participated. All available board-certified cardiologists holding a position at Levanger Hospital participated as reference examiners. The participants were chosen irrespective of personal motivation and previous experience with
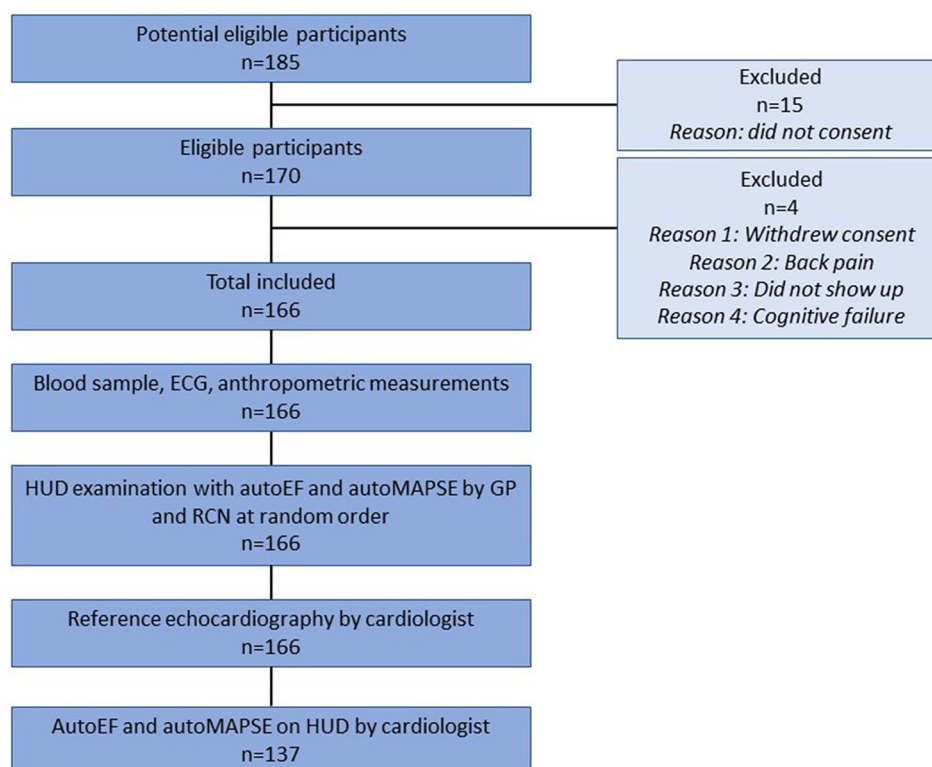

Figure 1 Patient flow through the study. Illustration of the patient inclusion, exclusion and order of examination throughout the study. In total, 160 patients were included. The first 29 patients were not examined with automatic HUD applications by the cardiologists due to logistic reasons. Abbreviations:AutoEF, automatic measurement of left ventricular ejection fraction by decision-support software; GP, general practitioner; HUD, handheld ultrasound device; autoMAPSE, automatic measurement of mitral annular plane systolic excursion by decision-support software; RCN, registered cardiac nurse.

Table 1 Image quality score

Score value	0	1	2	3	4	5	6
Four-chamber view		Two-chamber view, long axis view or others		Five-chamber view or posterior four-chamber view			Four-chamber view
LV alignment (misalignment)	≥45°		30°–44°		15°–29°		<15
Apical mispositioning			≥15 mm		<15 mm		None
Mitral annular assessment		Not judgeable	Poor	Fair	Good	Near excellent	Excellent
Number of visible endocardial segments		One	Two	Three	Four	Five	Six

This table represents five prespecified categories used to evaluate the image quality with the applied automatic decisionsupport software. Scores 0–6 were given for each category. The mean of each score per image stands for the overall quality of the image. LV, left ventricular.

ultrasound. As shown in figure 1, the study participants were examined at random order by one of five GPs and one of three RCNs blinded to their respective results. One of five cardiologists performed comprehensive echocardiography serving as the reference method, as well as apical HUD recordings for comparison. All users used the decision support software for automatic measurements of LV function by the HUDs; however, the first 29 patients were not examined by HUD by the cardiologists due to logistic reasons. There were no other examinations organised by the study.

Training and education

Details of training and education are comprehensively described previously.[8] In short, the GPs, RCNs and cardiologists represented the novice, intermediate and expert operators, respectively. The novices (n=6) underwent six days of one-to-one training supervised by one of two experienced cardiology fellows in addition to two evening lectures. They had access to private HUDs in their day-to-day practice for the whole training and inclusion period. Only one of the GPs had previous experience with ultrasound diagnostics (only seven examinations). One GP changed occupation and did not take part in the study leaving five GPs for the analyses. The intermediate group (n=3) were experienced in evaluation of pleural effusion, the inferior vena cava and limited ultrasound examinations of the heart.[5] Thus, they did not undergo systematic training, but were instructed on how to use the HUDs and initialise the automatic algorithms approximately four weeks prior to inclusion. The expert group consisting of in-house cardiologist experienced in echocardiography (n=5) were instructed at the day of inclusion on how to use the automatic measurements but did not receive any further instructions or training.

Ultrasound examinations

The study specific protocol has been described previously.[8] All participants underwent three HUD examinations using a VScan Extend (GE Ultrasound AS, Horten, Norway) in addition to the reference echocardiography (Vivid E9 or E95 scanner, GE Ultrasound AS). All HUD
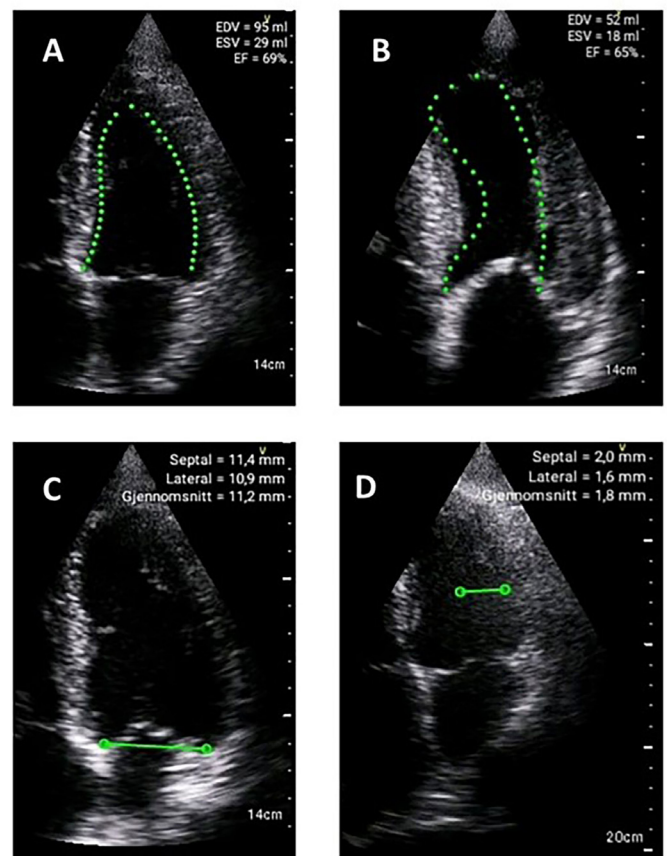
Figure 2 Examples of satisfactory and unsatisfactory decision support software results for ejection fraction and mitral annular plane systolic excursion. All images had an average score ≥5.6. (A) Satisfactory autoEF measurement with corresponding values. (B) Unsatisfactory autoEF measurement with values not reflecting true LV function. (C) Satisfactory autoMAPSE measurement with corresponding values. (D) Unsatisfactory autoMAPSE measurement with values not reflecting the true LV function. autoEF, automatic measurement of left ventricular ejection fraction by decision support software; autoMAPSE autoEF, automatic measurement of left ventricular mitral annular plane systolic excursion by decision-support software; LV, left ventricle.

Table 2 Baseline data, medications and comorbidities of the study population

Variable	
Age, years	73 (63–78)
Women, n (%)	78 (47)
Body mass index (kg/m^2)	28.7±5.3
Systolic blood pressure (mm Hg)	150±22
Diastolic blood pressure (mm Hg)	83±11
Glomerular filtration rate (mL/min)*	89 (67-89)
Haemoglobin (g/L)	144±15
N-terminal probrain natriuretic peptide (ng/L)	295 (66–864)
NYHA functional class, n (%)	
I	63 (37)
II	80 (47)
III	12 (7)
IV	1 (1)
Diuretics, n (%)	41 (25)
Beta blockers, n (%)	51 (31)
ACE inhibitor or angiotensin-receptor blocker, n (%)	32 (19)
Atrial fibrillation diagnosis, n (%)	49 (29)
Atrial fibrillation at day of inclusion, n (%)	40 (24)
COPD/asthma, n (%)	26 (16)
Diabetes mellitus type 2, n (%)	23 (14)
Coronary artery disease, n (%)	19 (11)

Normal distributed data are expressed as mean±SD, skewed data are presented as median (IQR) and proportions are n (%). Medications listed refer to the current use.
*Calculated by the Cockcroft-Gault equation.
COPD, chronic obstructive pulmonary disease; NYHA, New York Heart Association.

recordings by the three different user groups included apical four-chamber recordings with the addition of fully automated measurements of EF and MAPSE. All users ideally performed three separate recordings per automatic algorithm, a total of six four-chamber recordings per patient. The decision support software (autoEF or autoMAPSE) was then applied for fully automated measurements after each recording, and the analysed recordings were stored on the HUD. Subsequently, the recordings automatically transferred and stored on the Tricefy secure cloud-based server (Trice Imaging Inc, California, USA).

One of five cardiologists performed reference echocardiography on all patients in accordance with international recommendations for standardised echocardiographic examination.[9] LV endocardial borders were traced at end-diastole and end-systole in the four and two chamber views, and LV volumes and EF were calculated by the Simpson's biplane method. The mitral annular septal and lateral points in four-chamber views using motion mode measures MAPSE. All measurements were performed using EchoPAC SW Only, versions 202 and 203 (GE Ultrasound).

Automatic tools for quantification of LV function and image analysis

Details of the fully automated decision support software for quantification of LV function (autoEF and autoMAPSE) have been described elsewhere.[7 8 10] Shortly, the automatic measurements of LV end-diastolic volume (EDV), end-systolic volume and EF was done by the commercially available artificial intelligence aided LVivo EF software (DiA Imaging Analysis, Be'er Sheva, Israel). Fully automated tracing of the endocardial border in four-chamber recordings estimated LV volumes (figure 2). EF was calculated from the LV volume estimates based on the traces. The fully automated autoMAPSE software tracked the septal and lateral points of the mitral annulus in four-chamber recordings, and MAPSE was calculated as the average displacement of the septal and lateral points (figure 2).[7]

Image quality assessment

Image quality was evaluated by an external cardiologist experienced in echocardiography blinded to details of the operators and patients. All HUD recordings including either automatic desision-support software was reviewed. The image quality was scored by evaluating five prespecified categories, and the mean of the scores represents the averaged image quality score (table 1). Additionally, the external reviewer evaluated whether the automatic measurements were recommended for clinical use based on: (1) the image quality scores, (2) the quality of the tracking of the endocardial border for autoEF or the mitral annular points for autoMAPSE, respectively, and (3) the performance and numerical output of the autoEF and autoMAPSE algorithms. This was scored as following: (1) discard measurement (not for clinical use); (2) accept, but needs adjustment of the result due to suboptimal performance of the automatic software; (3) accept as it is.

During preliminary analyses by our group, detection of a system error in the autoEF software initiated a software revision by the vendor (LVivo EF, DiA Imaging Analysis, Be'er Sheva, Israel). The first 103 patients were examined with the first version of the autoEF software (version 1), and the following 63 patients were examined with the revised software (version 2).

Other data

Anthropometric measurements (body weight (kg), body height (cm) and blood pressure (mm Hg)) were measured, and New York Heart Association functional classification was scored by nurses the day of inclusion. Blood samples at the day of inclusion were analysed at the in-hospital accredited laboratory.

Statistics

Continuous variables were expressed as mean and SD or as median and IQR as appropriate. Normality was

Table 3 Pictures analysed per automatic function, per user and the average scores per parameter

	Novice	Intermediate	Expert
Automatic measurement of left ventricular ejection fraction			
Images, n	403	445	360
Total average score	3.7±0.9	4.2±0.9	4.8±0.8
Four-chamber view	3.9±1.7	4.7±1.7	5.1±1.5
LV alignment	4.7±1.3	5.5±1.0	5.6±0.8
Apical mispositioning	3.6±1.5	3.9±1.5	4.5±1.4
Mitral annular assessment	2.8±1.0	3.1±1.1	3.5±1.0
Number of visible LV endocardial segments	3.5±1.3	3.8±1.2	4.2±1.2
Automatic measurement of mitral annular plane systolic excursion			
Images, n	476	470	389
Total average score	3.2±1.9	3.7±0.9	4.4±0.8
Four chamber view	3.3±1.6	4.1±1.8	4.9±1.5
LV alignment	4.4±1.5	5.1±1.2	5.5±0.9
Apical mispositioning	3.0±1.3	3.3±1.4	4.0±1.4
Mitral annular assessment	2.5±0.9	2.9±1.0	3.5±1.0
Number of visible LV endocardial segments	2.8±1.3	3.2±1.3	3.9±1.3

The table illustrates the number and image quality score of the images by applications and users for the five prespecified parameters as well as the average score. All scores are given as mean±SD.
LV, left ventricular.

evaluated by histograms and Q–Q plots. Categorical variables were presented as frequencies and proportions. Student's t-test and Wilcoxon test were used for comparison of groups as appropriate. Proportions were compared using the χ^2 test and Fisher's exact test as appropriate. McNemar's test was used to compare paired nominal data. Repeated measure analysis of variance with post hoc Bonferroni correction was used to analyse variance in the groups. The influence of the image quality parameters with performance of the automatic applications, as well as patients' characteristics, was evaluated by logistic regression and general linear models as appropriate. The importance of image quality for feasibility and agreement with reference was first evaluated on the whole dataset of images from all three users and within the three user groups. The agreement with reference was assessed at the level of all available automatic measurements. The importance of the different image quality category for the within-patient reliability of the automatic applications was evaluated using the maximum difference in measurements of autoEF and autoMAPSE and the maximum difference in image quality scores. Analyses were performed in the whole dataset and within user groups. A p value <0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics, V.28 (SPSS Inc).

Initial calculations of sample size were 104 patients estimated by Sample Power (SPSS, Inc) based on diagnostic performance; however, in such a small population significant pathology would be scarce. Therefore, to account for likely low rate of pathological findings, the sample was expanded to 150 patients. Preliminary analyses revealed an error in the autoEF algorithm initiation a software upgrade, so recalculations of sample power led to an increase the population to 170 patients. No additional power analysis was performed in this study. The planned number of inclusions exceeds the number of participants needed for reliable evaluation of feasibility, reliability and agreement with paired analyses.

RESULTS

Study population

In total 185 patients with suspected heart failure were invited to take part, 15 did not consent, 1 withdrew consent, 1 could not complete the examinations due to back pain, 1 did not show up and 1 was excluded due to cognitive failure. In total, 166 participants were included in the analyses (figure 1). Population baseline characteristics are previously published but are shown in table 2.[11] Almost half the population was female, and mean age±SD was 70±13 years. Most patients were overweight with mean BMI±SD of 28.7±5.3 and a substantial proportion presented with atrial fibrillation (24%). Furthermore, chronic obstructive pulmonary disease (COPD) was present in 16% of patients.

Feasibility and image quality

In total, 2543 images were scored for assessment of image quality (table 3). Figure 3 shows that image quality score was consistently lower for novices versus intermediate experienced versus experts for both modalities. The image quality score was highest for the *LV alignment* and lowest for *mitral annular assessment*, with consistent findings across user groups and methods.

Feasibility was higher with higher image quality score. In univariate logistic regression analyses including all recordings irrespective of the user group, both the average image quality score and the five prespecified scores were associated with the feasibility of both automatic applications (all p<0.001). In multivariate analyses including all five image quality score categories, we found that all were significantly associated with the feasibility for autoEF (all p<0.001, except *four-chamber view* (p=0.02)). For autoMAPSE, *apical misposition* (p=0.94) and *number of visible LV endocardial segments* (p=0.06) were not significantly associated with the feasibility, while the other categories were (*four-chamber view*, p=0.046, others p<0.001).

Table 4 shows that image quality was more important for the feasibility of autoMAPSE than autoEF. Additionally, there was a gradient in adjusted R^2 ranging from 41% within novices to 22% within experts with respect to
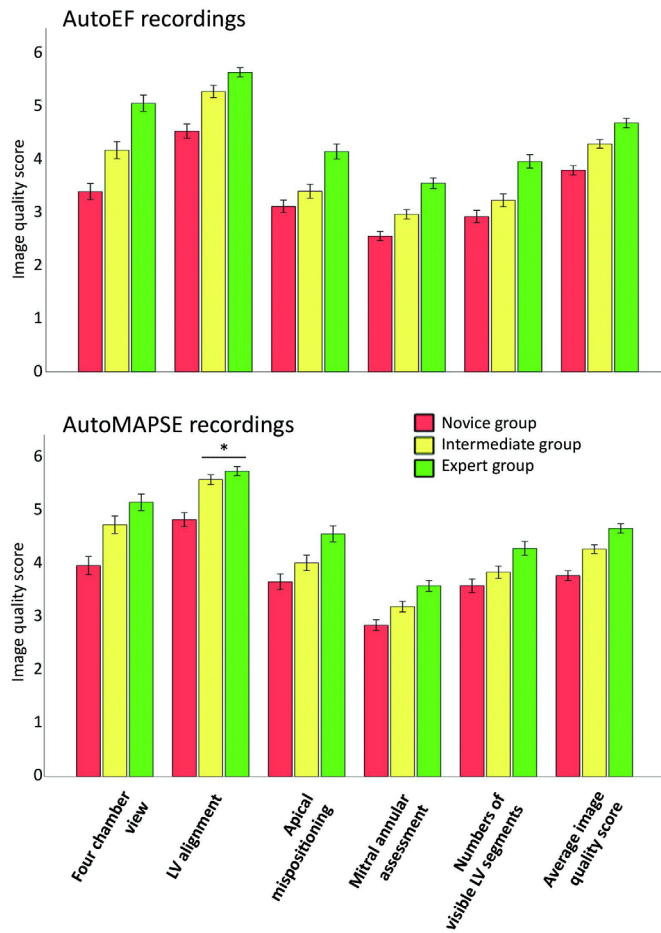
Figure 3 Image quality score parameters in handheld ultrasound recordings for automatic assessment of ejection fraction and mitral annular plane systolic excursion by user groups. figure 3 shows the image scores by user groups in recordings for automatic assessment of ejection fraction (autoEF) and mitral annular plane systolic excursion (autoMAPSE). All image scores were significantly different across user groups (p≤0.01) except for *LV alignment* between intermediate and expert users in autoEF recordings (*). autoEF, automatic measurement of left ventricular ejection fraction by decision support software; LV, left ventricle.

Table 4 The importance of image quality score components for feasibility of automatic assessment of ejection fraction and mitral annular plane systolic excursion

	Novice	Intermediate	Expert
All (five) image quality parameters	EF 20%/MM 41%	EF 18%/MM 37%	EF 24%/MM 22%
Four-chamber view	MM 13%		
LV alignment	EF 13%/MM 17%	EF 8%/MM 10%	
Apical mispositioning			EF 2%
Mitral annular assessment	EF 10%/MM 33%	MM 32%	MM 20%
Number of visible LV endocardial segments		EF 13%	EF 14%

The table shows the proportion of explained variance (adjusted R^2) for feasibility of automatic assessment by autoEF and autoMAPSE according to user groups, respectively. Each of the prespecified image score parameters were evaluated in univariate log-linear regression analyses, and parameters significantly associated with feasibility are shown. Subsequently, the five prespecified image quality scores were included in a multivariate regression analysis, where all shown data were still significantly associated with feasibility with p<0.001, except for the four-chamber view in novices (p=0.02).
AutoEF, Automatic measurement of ejection fraction; AutoMAPSE, Automatic measurement of mitral annular plane systolic excursion; EF, ejection fraction; LV, left ventricular; MAPSE, mitral annular plane systolic excursion; MM, motion mode.

Image quality was not significantly associated with known hypertension or COPD (R^2<1%, p>0.09).

In analyses comparing the importance of image quality for the feasibility of the different autoEF software versions, minor differences were revealed. The adjusted R^2 for version 1 was 23% for novices, 19% for the intermediate group and 27% for experts, with corresponding R^2 of 32%, 19% and 30% for version 2.

Agreement of HUD recordings with reference and image quality

Figure 4 shows that the image quality of the HUD recordings was not important for the agreement of the automatic decision support software measurements compared with reference. Image quality of the HUD recordings explained only 2% of the variability (R^2=2%) between the autoEF and reference measurements in the whole dataset. In analyses within user groups, the findings were similar (R^2=1% for all three user groups). Furthermore, the associations of less underestimation by the decision support software on HUDs with better image quality were only significant for the novices and the intermediate group. Similarly, image quality of the HUD recordings explained only 7% of the variability between the autoMAPSE and reference measurements in the whole dataset. In analyses within user groups, we found a gradient in the explained variance ranging from 7% for novices to only 1% for experts, however still significant across user groups (p<0.05 for all except autoEF in expert group p=0.056).

Reliability of decision support software measurements on HUDs and image quality

Figure 5 illustrates the within-patient differences for repeated measurements of EF, EDV and MAPSE by the decision support software according to within-patient differences in image quality in the whole dataset. In analyses combining all user groups, there were significant
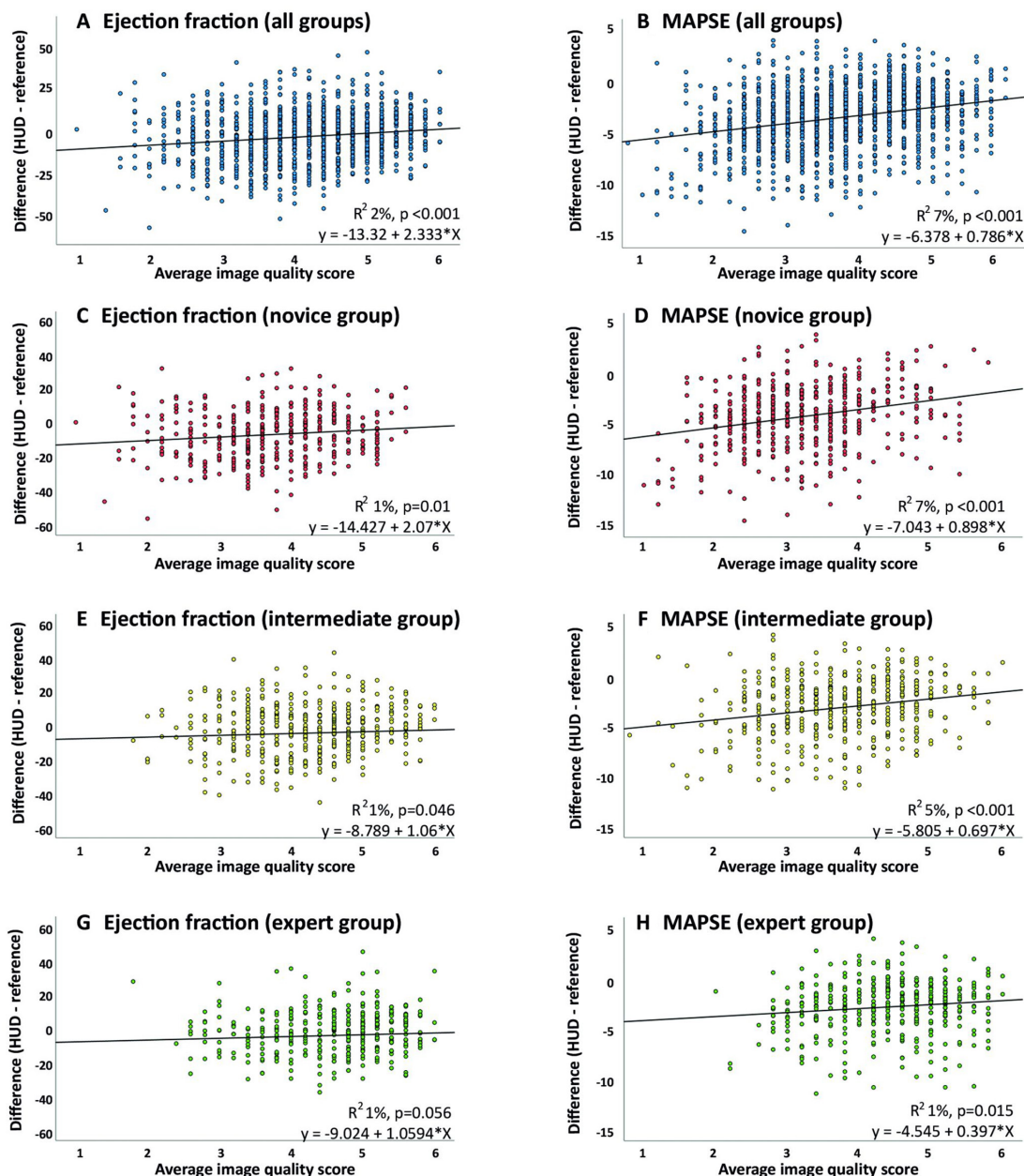
the feasibility of autoMAPSE, while no gradient was seen across user groups for autoEF. Among the image score categories, the *numbers of visible LV endocardial segments* were the most important predictor for the feasibility of autoEF for the two most experienced groups but not for novices. Correspondingly, *mitral annular assessment* explained most of the variability related to image quality for autoMAPSE across user groups.

The averaged image quality score was weakly associated with body mass index and systolic blood pressure for both autoEF and autoMAPSE (R^2≤4%, p≤0.04) when analysed in the whole dataset. Systolic blood pressure was not associated with image quality in experts (p≥0.42), while BMI showed stronger associations with image quality in experts (R^2 12% and 9% for autoEF and autoMAPSE, both p<0.001). In novices and the intermediate group, the associations with systolic blood pressure were very weak (R^2≤2%, p≤0.06).

Figure 4 Agreement between measurements of automatic ejection fraction and MAPSE by different operators using handheld ultrasound and reference echocardiography according to image quality. Scatter plots with applied 'line of best fit' of the difference in automatic ejection fraction and MAPSE measurements by HUDs and reference according to user groups by average image quality score. Panels A–H show measurement and user group. Explained variance of the difference between measurements are shown by the R^2, and the equation for the best fitted line (y-axis) is showed by image quality score (x-axis). MAPSE, mitral annular plane systolic excursion; HUD, handheld ultrasound device.

associations of lower reliability with larger within-patient variability in image quality of the repeated recordings (all $p \leq 0.005$). In user group specific analyses, the reliability was not significantly associated with image quality for neither of the three specified measurements in experts (all $p \geq 0.16$). For the novices and intermediate group, the reliability was significantly associated with within-patient differences in image quality for decision support software measurements of EF, but not for EDV ($p \geq 0.051$) or MAPSE ($p \geq 0.12$).

Discussion

This study evaluated the influence of operators' experience and image quality for fully automatic decision-support software measurements of LV EF, EDV and MAPSE by HUDs in three user groups with varying experience. Blinded evaluation of 2543 four chamber HUD recordings by novices, intermediate experienced users and experts showed that image quality was significantly associated with the feasibility of the decision support software measurements. Image quality was more closely
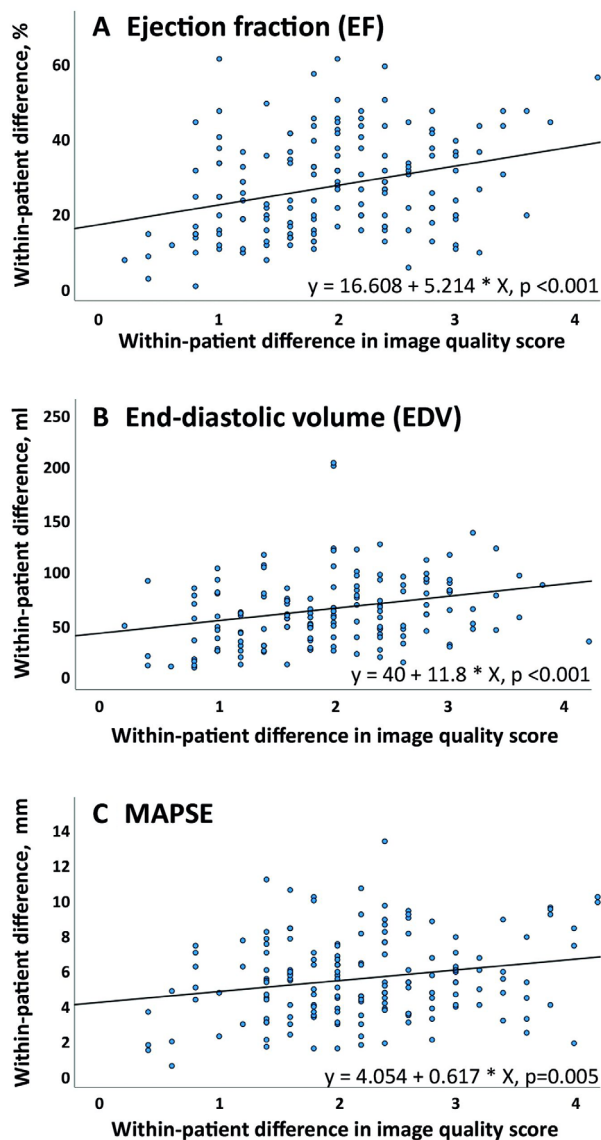
Figure 5 Within-patient differences in automatic measurements of ejection fraction, end-diastolic volume and MAPSE by handheld ultrasound devices plotted against the within-patient differences in image quality. The maximum within-patient difference for automatic HUD measurements of EF (A), EDV (B) and MAPSE (C) is plotted against the maximum within-patient difference in image quality score. Lines of best fit have been applied by linear regression calculation. MAPSE, mitral annular plane systolic excursion.

related to feasibility in the less experienced user groups and explained 18%–24% of the variability in feasibility of autoEF and volumes, and 22%–41% of the variability in the feasibility of autoMAPSE, respectively. Of five prespecified image quality categories, the *number of visible LV endocardial segments* was most important for autoEF and volumes, while *mitral annular assessment* was most important for autoMAPSE. Contradictory, the agreement of the automatic decision support software measurements was less dependent on image quality (adjusted $R^2 \leq 7\%$). Furthermore, image quality did not explain the low test–retest reliability of the decision support software

measurements. In user-specific analyses, the reliability of the decision support software measurements was significantly associated with image quality only for the less experienced user groups for autoEF measurements.

Population

The finding of elevated blood pressure and body mass index in significant proportions are expected as they represent relevant risk factors for HF, and the population studied was referred to hospital for evaluation of suspected HF. Similarly, since the population includes both healthy and diseased individuals, the distribution of relevant patient characteristics are wider compared with more strictly selected samples.[5 7 12] The presented associations of image quality with body mass index and systolic blood pressure do not seem to be of clinical importance with respect to the study aims.

Decision support software and image quality. Until recently, evaluation of LV function on HUDs has been done by visual assessment ('eyeballing') only, which has several limitations.[13] Easy to perform focused cardiac ultrasound performed by inexperienced users on HUDs is feasible and has showed promising results.[4 12 14] Image quality is essential and a major challenge within all ultrasound diagnostics. As overweight, atrial fibrillation and COPD were common in the studied population this challenges the image quality of the ultrasound recordings. As shown in this study the image quality was closely related to the experience of the users, even though the body mass index and systolic blood pressure were of importance as well.

The feasibility of both automatic decision software was higher when image quality score was higher. As shown, image quality influenced the feasibility of the automatic measurements more for the novices, and intermediate group, compared with the experts. This is related to less variation in image quality score for the experts and that the performance of the decision support software was not solely dependent on relevant image quality. The corresponding explained variance for the feasibility of autoMAPSE was nearly twice the explained variance for autoEF, indicating that the feasibility of autoMAPSE was closer associated with image quality. The *number of visible LV endocardial segments* category explained the majority of variance in feasibility for autoEF, and similarly the *mitral annular assessment* was most important for the feasibility of autoMAPSE. This finding is in line with clinical experience on echocardiographic requirements for EF and MAPSE. The finding of less influence of image quality for autoEF compared to autoMAPSE may be due to technological characteristics of the software. The autoEF software is assisted by artificial intelligence,[10] and it may be hypothesised that the training of the algorithm was not optimal for the HUD recordings used in this study. Second, the autoMAPSE software used grayscale images only, while the robustness of MAPSE is commonly shown for methods using tissue Doppler.[15]

Agreement and reliability

In a recent publication, we showed that the coefficient of repeatability for the presented automatic decision support software ranged 19%–24% (*reference article is not yet published but is currently in for review*). This is significantly higher than shown by experts using high-end ultrasound equipment[16], and a recent publication showing limits of agreement 14.5% using artificial intelligence assisted decision support software for assessment of EF by a novel HUD.[6] In the latter study, image quality evaluated by the number of LV walls where the endocardial border was not clearly identifiable in end-diastole did not significantly influence the agreement with reference.[6] By exploring the unacceptably high variability of the automatic desicion-support software presented in this study, only a minor part of the low agreement was explained by image quality. As shown by figure 4, both automatic decision-support software underestimated EF and MAPSE more compared to reference when image quality was low. Importantly, even for the experts where image quality overall was good, the agreement with reference was below recommendation for clinical use, and only 1% of the variation compared with reference was explained by image quality. Thus, this adds in the disfavour of clinical implementation of the presented automatic decision support software for LV evaluation on HUDs.

Similarly, for autoMAPSE, there was a linear relation of more underestimation compared with reference when image quality was low for all user groups. Still, the variability was too high even when image quality was good, and overall, only 7% of the variability compared with reference was due to image quality.

Adding to the low agreement of the automatic measurements by HUDs, we found that only a minor part of test–retest reliability was caused by differences in image quality. For the experts, we have recently published moderate to good intrarater intraclass correlation (0.72 for autoEF and 0.83 for autoMAPSE), and within-patient differences in image quality did not explain the modest reliability (p=0.16 for EF, p=0.45 for EDV and p=0.99 for MAPSE) (*reference article is not yet published but is currently in for review*). To our knowledge, image quality of repeated recordings and its relevance for the reliability of automatic decision-support software measurements of LV function has not been evaluated on HUDs previously. Figure 5 shows the importance of image quality for the reliability within patients, but these associations were not always present when performing the analyses per user groups. Importantly, in the experts' recordings, we found no signs that higher image quality improved the reliability of the automated decision-support software. This shows the inconsistency of the automatic measurements, indicating image quality alone not to be sufficient for reliable performance of the automatic decision-support software. Two other studies have evaluated the agreement of automatic evaluation of LV EF by HUDs.[6 10] However, direct comparison is difficult as the published data on image quality characteristics were scarce in these studies and both included only one experienced operator each.

Until recently no decision support software for evaluation of LV function evaluation has been available on HUDs. Automatic decision-support software for estimation of EF performed by experts has showed promising results in recent publications.[6 10] Furthermore, in a previous publication from our group, we showed a slight underestimation of autoMAPSE compared with reference.[7] However, experts do not usually seek or require decision support. Differences in the studied populations in LV function, arrhythmias and body composition may partly explain the differences between the studies. Further, we evaluated the two versions of the automatic decision-support software for EF calculations but to be consistent with the planned study aims, we did not reanalyse the patients analysed by the first software version. In the future, additional refinement of decision support software based on better training of the algorithms and artificial intelligence may improve the software. More advanced decision support software including deformation analyses will also be available for HUDs.

Strengths and limitations

The main strength of this study is the comprehensive blinded analyses of five distinct categories of image quality and the performance of the decision support software. Another strength is that the recruitment of inexperienced operators was based on positions in the community healthcare system and not based on motivation for participation. Furthermore, the three groups of operators (in total 13 different users) had different experience ranging from no previous experience to level III experience according to the American Society of Echocardiography.[17] However, with respect to reduce potential bias related to the user specific results, even larger groups of operators would have been preferred. The most important limitation is that we only reviewed images being able to run the decision support software. Thus, cases where the cardiac cycle or image quality did not allow for the applications to run were consistently excluded from the image quality analyses. This may influence the results between the user groups, as less recordings were able to run the decision support software among the less experienced user groups. Even though the cardiologists did not perform HUD examinations on the first 29 participants, the findings across user groups were consistent also in analyses of subgroups (data not shown).

CONCLUSIONS

Image quality was important for the feasibility of decision support software for automatic analyses of left ventricular ejection fraction, volumes and mitral annular plane systolic excursion by novices, intermediate experienced and expert groups using HUDs in a population with suspected heart failure. However, neither the low agreement with reference nor the modest within-patient

reliability are explained solely by image quality. Further refinement of the decision support software is warranted before implementing these into everyday practice for non-expert users of HUDs.

ORCID iDs
Anna Katarina Hjorth-Hansen http://orcid.org/0000-0003-4626-2113
Havard Dalen http://orcid.org/0000-0003-1192-3663

REFERENCES

1 Ponikowski P, Voors AA, Anker SD, *et al*. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* 2016;2016:2129–200.

2 Hu K, Liu D, Niemann M, *et al*. Methods for assessment of left ventricular systolic function in technically difficult patients with poor imaging quality. *J Am Soc Echocardiogr* 2013;26:105–13.

3 Popescu BA, Stefanidis A, Fox KF, *et al*. Training, competence, and quality improvement in echocardiography: the European association of cardiovascular imaging recommendations: update 2020. *Eur Heart J Cardiovasc Imaging* 2020;21:1305–19.

4 Cardim N, Dalen H, Voigt J-U, *et al*. The use of handheld ultrasound devices: a position statement of the European association of cardiovascular imaging (2018 update). *Eur Heart J Cardiovasc Imaging* 2019;20:245–52.

5 Hjorth-Hansen AK, Andersen GN, Graven T, *et al*. Feasibility and accuracy of tele-echocardiography, with examinations by nurses and interpretation by an expert via telemedicine, in an outpatient heart failure clinic. *J Ultrasound Med* 2020;39:2313–23.

6 Papadopoulou S-L, Sachpekidis V, Kantartzi V, *et al*. Clinical validation of an artificial intelligence-assisted algorithm for automated quantification of left ventricular ejection fraction in real time by a novel handheld ultrasound device. *EHJDH* 2022;3:29–37.

7 Magelssen MI, Palmer CL, Hjorth-Hansen A, *et al*. Feasibility and reliability of automatic quantitative analyses of mitral annular plane systolic excursion by handheld ultrasound devices: a pilot study. *J Ultrasound Med* 2021;40:341–50.

8 Magelssen MI, Hjorth-Hansen AK, Andersen GN, *et al*. The clinical influence of hand-held ultrasound examinations by general practitioners in patients with suspected heart failure supported by tools for automatic quantification and telemedicine. *Eur Heart J Cardiovasc Imaging* 2022;23:2022.

9 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:e14:1–39.

10 Filipiak-Strzecka D, Kasprzak JD, Wejner-Mik P, *et al*. Artificial Intelligence-Powered measurement of left ventricular ejection fraction using a handheld ultrasound device. *Ultrasound Med Biol* 2021;47:1120–5.

11 Hjorth-Hansen AK, Magelssen MI, Andersen GN, *et al*. Real-Time automatic quantification of left ventricular function by hand-held ultrasound devices in patients with suspected heart failure: a feasibility study of a diagnostic test with data from general practitioners, nurses and cardiologists. *BMJ Open* 2022;12:e063793.

12 Panoulas VF, Daigeler A-L, Malaweera ASN, *et al*. Pocket-size hand-held cardiac ultrasound as an adjunct to clinical examination in the hands of medical students and junior doctors. *Eur Heart J Cardiovasc Imaging* 2013;14:323–30.

13 Marwick TH. Ejection fraction pros and cons: JACC state-of-the-art review. *J Am Coll Cardiol* 2018;72:2360–79.

14 Gundersen GH, Norekval TM, Haug HH, *et al*. Adding point of care ultrasound to assess volume status in heart failure patients in a nurse-led outpatient clinic. A randomised study. *Heart* 2016;102:29–34.

15 Willenheimer R, Cline C, Erhardt L, *et al*. Left ventricular atrioventricular plane displacement: an echocardiographic technique for rapid assessment of prognosis in heart failure. *Heart* 1997;78:230–6.

16 Baron T, Berglund L, Hedin E-M, *et al*. Test-retest reliability of new and conventional echocardiographic parameters of left ventricular systolic function. *Clin Res Cardiol* 2019;108:355–65.

17 Wiegers SE, Ryan T, Arrighi JA, *et al*. 2019 ACC/AHA/ASE advanced training statement on echocardiography (revision of the 2003 ACC/AHA clinical competence statement on echocardiography): a report of the ACC competency management Committee. *Circ Cardiovasc Imaging* 2019;2019:12.