BMJ Open

Feasibility and reproducibility of real-time automatic quantification of left ventricular function by hand-held ultrasound devices in patients with suspected heart failure: A diagnostic accuracy study with data from general practitioners, nurses, and cardiologists

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-063793
Article Type:	Original research
Date Submitted by the Author:	09-May-2022
Complete List of Authors:	Hjorth-Hansen, Anna; Helse Nord-Trøndelag HF, Internal medicine; Norges teknisk-naturvitenskapelige universitet, MI Lab and Department of Circulation and Medical Imaging Magelssen, Malgorzata; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Andersen, Garrett; Levanger Hospital, Internal Medicine Graven, Torbjørn; Levanger Hospital, Internal Medicine Kleinau, Jens; Levanger Hospital, Internal Medicine Landstad, Bodil; Levanger Hospital, Department of Research Løvstakken, Lasse; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Skjetne, Kyrre; Levanger Hospital, Internal Medicine Mjølstad, Ole; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Dalen, Havard; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging
Keywords:	Heart failure < CARDIOLOGY, Echocardiography < CARDIOLOGY, Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS, Cardiovascular imaging < RADIOLOGY & IMAGING, Ultrasound < RADIOLOGY & IMAGING

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Feasibility and reproducibility of real-time automatic quantification of left ventricular function by hand-held ultrasound devices in patients with suspected heart failure: A diagnostic accuracy study with data from general practitioners, nurses, and cardiologists

Hjorth-Hansen Anna Katarina M.D[1,2]

Magelssen Malgorzata Izabela M.D*[1,3]

Andersen Garrett Newton M.D, Ph.D.[2]

Graven Torbjørn M.D[2]

Kleinau Jens Olaf M.D[2]

Landstad Bodil J, Ph.D.[4,5]

Løvstakken Lasse MSc, Ph.D.[1]

Skjetne Kyrre M.D[2]

Mjølstad Ole Christian M.D, Ph.D.[1,3]

Dalen Håvard M.D, Ph.D. [1,2,3]

[1]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

[2]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway.

[3]Clinic of Cardiology, St. Olavs University Hospital, Trondheim, Norway.

[4]Department of Health Sciences, Mid Sweden University, Sweden.

[5]Unit of Research, Education and Development, Östersund Hospital, Östersund, Sweden.

Address of correspondence

Anna K Hjorth-Hansen

Annakh@ntnu.no

Department of Circulation and Medical Imaging,

Norwegian University of Science and Technology,

Box 8905,

7491 Trondheim,

Norway

Abstract

Objectives: To evaluate the feasibility and reproducibility of hand-held ultrasound (HUD) examinations with real-time automatic decision-making software for ejection fraction (autoEF) and mitral annular plane excursion (autoMAPSE) by novices (general practitioners (GPs)), intermediate- (registered cardiac nurses (RCNs)) and expert users (cardiologists), respectively, compared to reference echocardiography by cardiologists.

Design: Diagnostic accuracy study.

Setting and participants: 166 patients with suspected heart failure underwent HUD examinations by five novices and three intermediate-skilled users including automatic measurements, and five experts performed HUD examinations with automatic measurements. HUD results were compared to a reference echocardiography. A blinded cardiologist scored the HUD recordings with automatic measurements as 1) discard, 2) accept, but adjust the measurement, or 3) accept the measurement as it is.

Primary outcome measure: The feasibility of automatic decision-making software for quantification of left ventricular function.

Results: The overall feasibility for autoMAPSE or autoEF was >80%. The proportion of images judged feasible (score of ≥2) was lowest for novices and highest for experts for both autoEF and autoMAPSE (p≤0.001). Large coefficients of variation and wide coefficients of repeatability indicate moderate agreement. The corresponding intraclass correlations (ICC) were moderate to good (ICC 0.51-0.85) for intra-, and poor (ICC 0.35-0.51) for inter-rater analyses. The modest to poor agreement and reliability were not explained by the experience and competence of the users only.

Conclusion: Novices, intermediate and expert users were able to record four-chamber views for automatic assessment of autoEF and autoMAPSE using hand-held ultrasound devices. However, the modest agreement and reliability highlight the need for more reliable methods before implementing into clinical practice.

Keywords: Heart failure, ejection fraction, mitral annular plane systolic excursion, agreement, reliability, diagnostic

Strengths and limitations:

- To our knowledge, no study has evaluated automatic real-time quantification of left ventricular function on hand-held ultrasound devices by inexperienced users. The three user groups had different levels of experience, ranging from no previous experience to American Society of Echocardiography level III.
- The inexperienced operators were recruited by their role in the municipality and not based on motivation for attending the study.
- There is no gold standard for evaluation of left ventricular function and echocardiographic measurements by experienced cardiologists were used as reference.
- An error was detected in the first software version of the autoEF decision-support software, and this experience may have affected the results for the revised software as well.
- The study sample is expected to provide adequate power for analyses.

Introduction

Heart failure (HF) is a severe condition with poor prognosis and reduced quality of life which constitutes a burden on the health care system with high costs (1) and 26 million patients affected worldwide (2). Echocardiography is the cornerstone imaging modality for diagnostics and patient follow-up. Correct diagnosing can be challenging. It is shown that (in-training) cardiology fellows inaccurately interpret echocardiograms (3) and delay in diagnosis may be present in up to 40% (4).

Estimation of left ventricular (LV) ejection fraction (EF) is required for classification and treatment of HF (5). Another robust and easily obtainable measure of LV function is mitral annular plane systolic excursion (MAPSE) which is quite sensitive for detection of LV dysfunction (6, 7), even when EF is preserved. Semi-automatic quantification of LV EF has been available for some time, but automatic quantification of MAPSE is scarcely available (8).

Hand-held ultrasound devices (HUD) has been widely implemented in the medical field over the last decade and increasingly used by non-experts (9). So far, quantification of LV size and function by HUDs relies on visual evaluation only (10). Several studies have shown high feasibility and reliability for inexperienced users performing simple tasks by HUDs (11-15). The experience and skill of the operator is essential for more advanced measures such as assessment of LV function (15, 16). Automatic measurement of LV EF (autoEF) from apical HUD recordings are now commercially available, and a novel method for real-time automatic measurement of MAPSE (autoMAPSE) is available on the GE Vscan Extend (GE Utrasound AS, Horten, Norway) for research purposes. This allow for real-time quantification of LV function by HUDs, and thus, it is a need to evaluate the feasibility and reproducibility in clinical scenarios by different users before implementing into clinical practise.

We aimed to evaluate the feasibility and reproducibility of HUD examinations including real-time autoEF and autoMAPSE performed by user of different levels of experience. Specifically, the novice, intermediate, and expert groups were represented by general practitioners, specialized cardiac nurses, and experienced cardiologists, respectively. Comprehensive echocardiography by experienced cardiologists served as reference.

Methods

Study design

Figure 1 indicates the flow of the study participants. The patients were examined by one of five general practitioners (GPs) and by one of three registered cardiac nurses (RCNs) at random order. GPs and RCNs, representing novice and intermediate level examiners, were blinded to each other's results. Reference echocardiography was performed by one of five cardiologists blinded to preceding examinations. An additional HUD examination was performed by the cardiologists (expert group). Due to logistic reasons, the first 29 patients were not examined by HUD by the cardiologist. No further follow-up or ultrasound examinations of the participants were performed during the study.

Participants

Patients referred to Levanger Hospital, Norway, with suspected HF were available for inclusion. Exclusion criteria were age <18 years, known HF and previous cardiac imaging within the last decade. Participants were included from June 2018 to June 2020. Inclusion was paused from March to June 2020 due to the COVID-19 pandemic. All participants gave their informed, written consent prior to inclusion. The study was performed in conformity with the Declaration of Helsinki.

Training and education of personnel

The conductors of the study had no influence on the selection of GPs for the study which were selected by the municipality administration based on the GPs' position in the two municipalities of Levanger and Verdal.

A total of six GPs underwent training in focused cardiac ultrasound by HUD. One dropped out due to change of occupation, and thus, five participated in the study. They underwent six in-hospital training days with one-to-one supervision by one of two residents experienced in focused cardiac ultrasound, in addition to two evening lectures provided by experts in diagnostic ultrasound and echocardiography. The GPs had the opportunity to use a personal HUD without supervision from the first day of training, but for no longer than three months prior to inclusion. None of them received additional training prior to study start. Upon direct request, no GP considered himself/herself underprepared to start inclusion. Only one of the six had performed focused ultrasound examinations prior to training (N=7), and thus, the group represents inexperienced users. They performed in total median (range) 46 (45-68) examinations prior to first inclusion, where median (range) 10 (9-20) were unsupervised and 36 (31-43) supervised, respectively.

Three RCNs with experience from a nurse-led outpatient HF clinic represent intermediate experienced users. They had experience in evaluation of pleural effusion, the inferior vena cava and clinical signs in HF patients. They had also participated in previous studies with limited ultrasound examinations of the heart (17). The RCNs had completed a total of median (range) 118 (74-221) limited echocardiographic examinations before patient inclusion, and thus, they did not undergo the same systematic training as the GPs. They were

instructed in how to use the HUD and initialize autoEF and autoMAPSE approximately 4 weeks prior to inclusion.

Five cardiologists experienced in echocardiography were only instructed in how to initialize the automatic tools on the HUDs and were not provided any additional training.

Test method

Each patient underwent three HUD examinations in addition to the reference imaging. All HUD examinations were performed by a Vscan Extend, and similarly, reference echocardiography by a Vivid E9 or E95 scanner (GE Ultrasound AS, Horten, Norway). All examinations were performed according to standard operating procedures and included 4-chamber recordings of the LV. The protocol for the GPs included parasternal long- and short-axis views, apical 4-chamber view, subcostal 4-chamber view, evaluation of the inferior vena cava and the pleural cavities. The recording of the inferior caval vein included both maximum and minimum dimension by including inspiration. Pleural cavities were assessed in in sitting position, and in case of pleural effusion craniocaudal images were recorded. RCNs recorded the same above-mentioned views, as well as apical 2-chamber and apical long-axis views, right ventricular focused 4-chamber view, and atrial focused recordings. Additionally, RCNs recorded colour Doppler images of the mitral, aortic, and tricuspid valve. Cardiologists recorded 4-chamber view only on the HUD, but the reference echocardiography was comprehensive (18).

For all HUD examinations live cine-loops of at least one cardiac cycle were recorded. The software for autoEF or autoMAPSE was launched by the operator and the automatic

analysed recordings were subsequently stored on the HUD. This was repeated aiming for 6 separate recordings for automatic analyses by autoEF (3 recordings) and autoMAPSE (3 recordings). All recorded views and analyses were stored and transferred without delay to a Tricefy® cloud-based server (Trice Imaging Inc., CA, USA).

Reference echocardiographic examinations were performed according to recommendations (18) in a separate room immediately after the examinations by the GPs and RCNs. All measurements reflect the average of at least three (five in the case of arrhythmia) cardiac cycles. Central methodology follows: All measurements were performed using EchoPAC SW Only, version 202 and 203 (GE Ultrasound). The LV endocardial borders were traced in end-diastole and end-systole in 4-chamber and 2-chamber view. LV volumes (end-diastolic and end-systolic) and EF was calculated based on the traces using the biplane Simpson's method. MAPSE was measured as the longitudinal displacement of the mitral annular septal and lateral points in reconstructed motion mode.

Details of the automatic tools for quantification of LV function and image analyses

Before storing of the loop of the 4-chamber view, the specific application (autoEF or autoMAPSE) was launched on the HUD. The automatic measurements of LV volumes and EF was done by the commercially available LVivo® app (DiA Imaging Analysis, Be'er Sheva, Israel). The app provides fully automatic edge detection and tracing of the endocardial border in standard apical 4-chamber views throughout the cardiac cycle. LV volume was estimated at end-diastole and end-systole and EF was calculated from the volume estimates. MAPSE was estimated by an automated algorithm tracking the mitral annular septal and lateral points using a LV model. Technical details of the method is described in a previous paper (19). Shortly, a Real-time Contour Tracking Library (RCTL) was used to process and

track the LV movement and images (GE, Vingmed, Norway) using a non-uniform rational B-spline model (20). The mitral annular septal and lateral points of the model were returned from RCTL. The array of points was evaluated to locate the maximum mitral annular plane displacement. MAPSE was calculated at the septal and lateral mitral annular points and as averaged values. For both autoEF and autoMAPSE the 4-chamber view recording with the overlay of the results from the automatic algorithm was stored as described above.

All HUD recordings were made available for blinded analyses by external cardiologists experienced in echocardiography. These cardiologists scored all recordings with the automatic measurement overlay as one of the following categories: 1) Discard (not for clinical use), 2) Accept, but adjust the result according to suboptimal performance, or 3) Accept the result as it is. The scoring took both the quality of the recordings and the performance of the application used into account. Thus, if the recording was not representative for a 4-chamber view the score was lower. The latter part of the scoring was based on identification and tracking of the endocardial border (autoEF), or mitral annular points (autoMAPSE) combined with the numerical output.

During the study we detected an error in the autoEF software, and thus, the LVivo app was revised by the vendor during the summer of 2019. In total 103 were analysed with the first version of the autoEF software and 63 patients with the revised software.

Other measurements

Blood samples were drawn the same day and analysed at the in-hospital accredited laboratory. Serum N-terminal pro-Brain Natriuretic Peptide (NT-pro-BNP), serum creatinine and estimated glomerular filtration rate (eGFR, calculated by the Cockcroft-Gault equation),

as well as electrolyte status (serum sodium and serum potassium) and haemoglobin (g/dL) were measured. New York Heart Association (NYHA) functional classification was scored by the nurses, and body weight (kg), body height (cm) and blood pressure (mmHg) were measured. Anthropometric measurements were rounded up to the nearest multiple of 1.

Patient and public involvement

Patients were not involved in setting the research question or the outcome measures. However, patient user groups were involved in planning of the study period as well as the way of informing the patients and the society of the study results.

Analyses

Continuous variables were expressed as mean and standard deviation (SD) or as median and interquartile range (IQR) as appropriate. Evaluation of normality was done by evaluation of histograms and normality plots. Categorical variables are presented as frequencies and proportions. Student's *t*-test and Wilcoxon test was used for comparison of groups when appropriate, ANOVA with post-hoc LSD correction were used to compare several groups. A study was judged as feasible if the user was able to acquire data with the fully automatic applications combined with that the blinded scoring by the cardiologist was that the recording and automatic measurement was accepted for clinical use. Proportions were compared using the Chi square test and Fisher's exact test when appropriate. Reliability of the measurements was evaluated by intraclass correlations (ICC), where values <0.5 were considered poor, 0.5-0.75 moderate, 0.75-0.9 good and >0.9 excellent (19). The intra-rater reliability was calculated by a two-way mixed model defined by absolute agreement in the

dataset of single measurements analysed by the automatic methods. The inter-rater reliability was calculated with a two-way random model defined by absolute agreement in the dataset of average measurements analysed by both the GPs, nurses and cardiologists by HUDs compared to reference. The agreement with reference echocardiography was evaluated by coefficients of variation, coefficient of repeatability indicating the minimal detectable change and Bland-Altman statistics. A p-value <0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics, version 27 (SPSS Inc, Chicago, IL, USA).

Results

Participants

Baseline characteristics shown in Table 1. In total, 185 patients were invited to participate, 170 were included and four were excluded (no show (1), cognitive failure (1), withdrew consent (2)). Shortly, 166 participants were included (47% women), mean ± SD age was 70 ± 13 years. NT-proBNP was above 125 ng/L in 101 (61%) with an overall mean of 705 ng/L. More than half the population (93 (55%)) was in NYHA class ≥II and had obesity or overweight (123 (74%)). Chronic pulmonary diseases were relatively rare (24 (15%)). Atrial fibrillation was known in 49 (29%) of the patients, and present at inclusion in 40 (23%).

Table 1. Baseline data, medications, and comorbidities of the study population

Variable	
Age, years	70 ± 13 (22-92)
Women, n (%)	78 (47)

Body mass index (kg/m^2)	28.7 ± 5.3
Systolic blood pressure (mmHg)	150 ± 22
Diastolic blood pressure (mmHg)	83 ± 11
Glomerular filtration rate (ml/min) *	67 ± 18 (15-125)
Haemoglobin (g/dL)	14.4 ± 1.5
N-terminal pro brain natriuretic peptide (ng/L)	705 ± 1219 (6-11309)
NYHA functional class	
I, n (%)	63 (37)
II, n (%)	80 (47)
III, n (%)	12 (7)
IV, n (%)	1 (1)
Diuretics, n (%)	41 (25)
Beta blockers, n (%)	51 (31)
Angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker, n (%)	32 (19)
Atrial fibrillation, n (%)	49 (29)
Chronic obstructive pulmonary disease/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 mean ± SD (interquartile range) and proportions are n (%). Medications listed refer to the current use. Abbreviations: *Calculated by the Cockcroft-Gault equation.

Test results

Feasibility

The novices were able to record at least one 4-chamber image with autoEF and autoMAPSE in 134 (80%) and 153 (92%) patients, respectively. The corresponding numbers for the intermediate group were 151 (90%) and 161 (96%), respectively (difference versus novices, both p<0.001). The experts were able to obtain the same measurements by the HUD for autoEF in 91% of the cases and autoMAPSE in 99% (difference versus the intermediate group, both p<0.001).

The proportion of images judged as feasible (score of ≥2) by the blinded cardiologist (Table 2) was lowest for novices, higher for the intermediate group and highest for experts for both autoEF and autoMAPSE (all p≤0.001). Overall, ≤53% of images with autoEF or autoMAPSE by novices were judged feasible, increasing to 84% and 85% for autoEF and autoMAPSE by experts, respectively. In analyses taking the two versions of the autoEF algorithm into account, the feasibility for autoEF improved after the update for all examiners ranging from 68% for novices to 91% for experts (Table 2). Only very few recordings with the automatic algorithm overlay were scored as 3: “Accept the result as it is”. In total, n (%) for autoEF and autoMAPSE were 7 (2%) and 23 (5%) for novices, 13 (3%) and 52 (11%) for the intermediate group and 25 (7%) and 67 (17%) for experts. The proportion of feasible recordings using autoEF was lower for the revised autoEF algorithm in novices and experts, and only 12 recordings in total were scored as 3: “Result accepted as is” for all users after revision.

The time used for the focused cardiac ultrasound examination was mean (SD) 18 (7) min for novices and 23 (7) min for the intermediate group. The time used for the six recordings with the automatic measurements was mean (SD) 4 min 34 sec (2min 20 sec) for novices, 3 min 21 sec (1 min 52 sec) in the intermediate group and 2 min 21 sec (1 min 19 sec) in experts, respectively.

Table 2. Feasibility of image recording and the use of automatic applications.

	Hand-held ultrasound operator		
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**
AutoEF, all patients	205/400 (51%)	296/442 (67%)	298/357 (84%)
AutoEF, first software version	100/246 (41%)	149/270 (55%)	148/193 (77%)
AutoEF, revised software version	105/154 (68%)	147/172 (85%)	150/164 (91%)
AutoMAPSE, all patients	248/471 (53%)	335/467 (72%)	333/391 (85%)

Data are presented as number/available recordings (%). All recordings with a feasibility score of ≥2 meaning accepted with or without need for adjustments. Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; GP, general practitioner; RCN, registered cardiac nurse.

Reproducibility

Table 3 shows the agreement of autoEF and autoMAPSE by the different users with reference. Shortly, the large coefficients of variability and large coefficients of repeatability for all three examiners indicate poor agreement between the automatic applications compared to reference. There was only a modest difference with respect to agreement between the operators. The minimal detectable change estimated from the coefficient of repeatability for autoEF and autoMAPSE ranged 24.2-21.5% and 5.0-4.1 mm, respectively. After the revision of the autoEF software, the minimal detectable change was somewhat improved but still approximately 20%.

Table 4 shows that intra-rater ICCs were moderate for all user groups with values <0.75 for all except for autoMAPSE by the intermediate group (0.85) and experts (ICC 0.83). The intra-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.51-0.72. The intra-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.70-0.85, respectively.

The inter-rater ICC was poor (≤0.51) for both automatic applications and all users. Inter-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.43-0.51. The inter-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.35-0.51, respectively.

Figures 2 show the Bland-Altman plots of the mean difference versus bias for all HUD recordings with automatic applications versus reference per operator. Similarly, Figure 3 is limited to images accepted (score 2 or 3) by the blinded cardiologist. Overall, the agreement was poor to moderate. We found no association of size of the measurement with agreement, but the limits of agreement were lower for the most experienced users (also shown in Table 3) and after excluding the images deemed too poor for clinical use (Figure 3).

Table 3. Mean values and the agreement of automatic hand-held ultrasound measurements of left ventricular function compared to reference.

	Hand-held ultrasound operator			
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**	**Reference echocardiography**
	Mean and agreement, autoEF (all recordings)			
Mean (SD), %*	51.7 (10.1)	52.9 (9.6)	53.3 (9.5)	53.4 (10.1)

Coefficient of variation, %	15.4	13.3	12.0	-
Coefficient of repeatability, %*	24.0	24.2	21.5	-
Mean and agreement, autoEF (first software version, n=107)				
Mean (SD), %*	52.6 (11.6)	54.2 (10.3)	55.0 (10.4)	53.5 (10.0)
Coefficient of variation, %	14.8	13.5	11.2	-
Coefficient of repeatability, %*	24.7	24.6	21.4	-
Mean and agreement, autoEF (revised software version, n=63)				
Mean (SD), %*	50.8 (8.4)	51.0 (8.3)	51.6 (8.1)	54.7 (9.6)
Coefficient of variation, %	16.0	13.1	12.9	-
Coefficient of repeatability, %*	20.6	20.6	19.8	-
Mean and agreement, autoMAPSE (all patients)				
Mean of septal and lateral position				
Mean (SD), mm	9.8 (2.4)	10.1 (2.6)	10.2 (2.5)	11.4 (2.9)
Coefficient of variation, %	24.3	20.5	18.9	-
Coefficient of repeatability, mm	5.0	4.8	4.1	-

Comprehensive echocardiography by experienced cardiologists used as reference. *%-points. Abbreviations: As described in Table 2.

Table 4. Intra- and inter-rater reliability of automatic measurements of left ventricular function by HUDs according to operators.

	HUD measurements by		
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**
Intra-rater intraclass correlation (ICC)			

AutoEF	0.58*	0.51	0.72
AutoMAPSE	0.70*	0.85	0.83
Inter-rater intraclass correlation (ICC)			
AutoEF	0.44	0.43	0.51
AutoMAPSE	0.35	0.44	0.51

Intra-rater intraclass correlation (ICC) calculated from single recordings per patient with automatic quantification of left ventricular function. Inter-rater intraclass correlation based on average values per patient and operator. *ICC of two repeated measures as only 38 patients had only three repeated measures of autoEF and 50 of autoMAPSE, respectively. Abbreviations: As described in Table 2.

Discussion

This study evaluated the feasibility and the reproducibility of real-time automatic quantification of LV function by HUDs by users of different levels of experience. The main findings were that the feasibility of the applications was acceptable, and the experienced users had the highest feasibility. Agreement with reference was poor to moderate, and even for the experts the agreement and reliability was barely within recommended ranges for clinical use.

The study population represents patients referred for cardiac examination to rule-in or rule-out HF in everyday practice. The novices underwent limited, but dedicated training. The intermediate group utilized focused cardiac ultrasound in in their clinical practice, and the experts were experienced with echocardiography and HUDs. The training of novices was in

line with comparable studies and present recommendations (10, 21). Most of the patients were overweight or obese with common comorbidities as atrial fibrillation and hypertension. Thus, both poor acoustics and atrial fibrillation (presented in 29%) could interfere with image acquisition and accuracy of the automatic measurements.

The overall feasibility was high for autoEF and autoMAPSE with >80% and >92% success rate for performance by all user groups when no quality assessment of the recorded image or performance of the applications was performed. The proportions were lowest for the novices and highest for the experts. The feasibility of the autoEF application significantly improved after revision. However, after blinded quality assessment by the cardiologist the feasibility was markedly impaired for both applications. The time consumption for the complete HUD examinations was mean 18-23 min for novices and the intermediate group, which we believe, is acceptable in the everyday practice in selected cases if the potential for clinical benefit is significant.

The intra- and inter-rater ICCs for novices and the intermediate group, were mainly lower than what would be recommended for clinical use (commonly used cut-off of 0.7). For experts the ICCs were somewhat higher, but compared with reference only 0.51, and in intra-rater analyses 0.72-0.83, respectively. Thus, we find that image quality and operator experience alone cannot explain the only moderate intra-operator reliability among the experienced cardiologists. Further studies must address how the next generation automatic analyses of LV function will perform.

The agreement was poor for automatic measurements of EF and MAPSE for all users, and only for experts the inter-rater ICC exceeded 0.5 (0.51). Even though the bias for autoEF was lower for the most experienced users, the agreement was poor to moderate for all

subgroups. In addition, limiting the analyses to recordings with automatic application overlay accepted by the blinded cardiologist. For autoMAPSE the underestimation compared to reference was consistent and this is known from previous studies by our group(19). This highlight that cut-off for pathology is not interchangeable between different methods. Suboptimal image acquisitions by inexperienced users partially explains the results, but importantly the agreement was suboptimal also in experts. This may indicate that the algorithms behind the applications need refinement before incorporation as a reliable tool in everyday clinical work independently of the skills of the users.

Conclusion

Inexperienced general practitioners, intermediate experienced registered cardiac nurses and expert cardiologists were able to perform automatic analyses of left ventricular function by automatic applications implemented in hand-held ultrasound devices. However, the different measures of the reproducibility of these automatic applications showed poor to moderate agreement with reference and a modest reliability. This study is a step in the right direction using novel technology to aid clinicians in diagnostic decision-making, methods that are more reliable are needed before large-scale implementation.

Acknowledgements

The Department of Circulation and Medical Imaging, Norwegian University of Science and Technology hosts a research collaboration between university, hospitals, and various vendors funded by the Norwegian Research Council. GE Ultrasound is one of these partners but had no role in planning or performing this study. We want to thank clinicians and other

employees at Nord-Trøndelag Hospital Trust for their support and for contributing to data collection in this research project.

Ethics approval statement

The study was approved by the regional committee for medical and health research ethics (REK 2017/2054) and registered in the ClinicalTrial.gov database (NCT03547076).

Author contributions

Hjorth-Hansen Anna Katarina M.D: Protocol description, data collection, data analyses, manuscript draft and revision

Magelssen Malgorzata Izabela M.D: Data collection, manuscript draft and revision

Andersen Garrett Newton M.D, Ph.D: Data acquisition, manuscript revision

Graven Torbjørn M.D: Data acquisition, manuscript revision

Kleinau Jens Olaf M.D: Data acquisition, manuscript revision

Skjetne Kyrre M.D: Data acquisition, manuscript revision

Landstad Bodil J, Ph.D: manuscript revision

Løvstakken Lasse MSc, Ph.D.: Software development, contribution to study design, manuscript revision

Mjølstad Ole Christian M.D, Ph.D: Data acquisition, manuscript revision

Dalen Håvard M.D, Ph.D: Study design, data acquisition, manuscript revision

Competing interests

GE Ultrasound provided the HUD devices for loan through a research contract with the project leader (HD), but GE had no role in performance of the study. AKHH, MIM, OCM, LL and HD hold positions in Centre for Innovative Ultrasound Solutions where GE Ultrasound is one of the industrial partners. LL acts as part-time consultant for GE Ultrasound.

Funding

The study was funded by grants from the European Interreg A initiatives (Norwegian-Swedish initiative), Research Council of Norway and Nord-Trøndelag Hospital Trust.

Data sharing statement

Not available.

References

1. Maggioni AP, Dahlstrom U, Filippatos G, Chioncel O, Leiro MC, Drozdz J, et al. EURObservational Research Programme: the Heart Failure Pilot Survey (ESC-HF Pilot). Eur J Heart Fail. 2010;12(10):1076-84.
2. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, et al. Heart failure: preventing disease and death worldwide. ESC Heart Fail. 2014;1(1):4-25.
3. Zhitny V, Iftekhar N, Alexander L, Ahsan C. Cardiology Fellow Diagnostic Accuracy and Data Interpretation Outcomes: A Review of the Current Literature . Vascular Health and Risk Management. 2020;Volume 16:429-35.
4. Bottle A, Kim D, Aylin P, Cowie MR, Majeed A, Hayhoe B. Routes to diagnosis of heart failure: observational study using linked data in England. Heart. 2018;104(7):600-5.
5. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 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;37(27):2129-200.
6. R W, C C, B I. Left ventricular atrioventricular plane displacement: an echocardiographic technique for rapid assessment of prognosis in heart failure. Heart. 1997(78):230-6.
7. Vinereanu D, Nicolaides E, Tweddel AC, Fraser AG. "Pure" diastolic dysfunction is associated with long-axis systolic dysfunction. Implications for the diagnosis and classification of heart failure. Eur J Heart Fail. 2005;7(5):820-8.
8. Snare SR, Mjolstad OC, Orderud F, Haugen BO, Torp H. Fast automatic measurement of mitral annulus excursion using a pocket-sized ultrasound system. Ultrasound Med Biol. 2011;37(4):617-31.
9. Díaz-Gómez JL, Ingelfinger JR, Mayo PH, Koenig SJ. Point-of-Care Ultrasonography. New England Journal of Medicine. 2021;385(17):1593-602.
10. Cardim N, Dalen H, Voigt JU, Ionescu A, Price S, Neskovic AN, 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(3):245-52.
11. Andersen GN, Graven T, Skjetne K, Mjolstad OC, Kleinau JO, Olsen O, et al. Diagnostic influence of routine point-of-care pocket-size ultrasound examinations performed by medical residents. J Ultrasound Med. 2015;34(4):627-36.
12. Andersen GN, Viset A, Mjolstad OC, Salvesen O, Dalen H, Haugen BO. Feasibility and accuracy of point-of-care pocket-size ultrasonography performed by medical students. Bmc Medical Education. 2014;14.
13. Gundersen GH, Norekval TM, Haug HH, Skjetne K, Kleinau JO, Graven T, 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(1):29-34.
14. Lipczyńska M, Szymański P, Klisiewicz A, Hoffman P. Hand-Carried Echocardiography in Heart Failure and Heart Failure Risk Population: A Community Based Prospective Study. Journal of the American Society of Echocardiography. 2011;24(2):125-31.
15. Panoulas VF, Daigeler AL, Malaweera AS, Lota AS, Baskaran D, Rahman S, 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(4):323-30.
16. Mjolstad OC, Andersen GN, Dalen H, Graven T, Skjetne K, Kleinau JO, et al. Feasibility and reliability of point-of-care pocket-size echocardiography performed by medical residents. Eur Heart J Cardiovasc Imaging. 2013;14(12):1195-202.
17. Hjorth-Hansen AK, Andersen GN, Graven T, Gundersen GH, Kleinau JO, Mjølstad OC, 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. Journal of Ultrasound in Medicine. 2020;39(12):2313-23.
18. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015;28(1):1-39 e14.
19. Magelssen MI, Palmer CL, Hjorth-Hansen A, Nilsen HO, Kiss G, Torp H, et al. Feasibility and Reliability of Automatic Quantitative Analyses of Mitral Annular Plane Systolic Excursion by Handheld Ultrasound Devices. Journal of Ultrasound in Medicine. 2020;40(2):341-50.
20. Storve S, Grue JF, Samstad S, Dalen H, Haugen BO, Torp H. Realtime Automatic Assessment of Cardiac Function in Echocardiography. IEEE Trans Ultrason Ferroelectr Freq Control. 2016;63(3):358-68.
21. Spencer KT, Kimura BJ, Korcarz CE, Pellikka PA, Rahko PS, Siegel RJ. Focused cardiac ultrasound: recommendations from the American Society of Echocardiography. J Am Soc Echocardiogr. 2013;26(6):567-81.

Figure legends.

Figure 1.

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

Figure 2.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography in the whole material (without exclusion of inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

Figure 3.

Bland-Altman plots illustrating agreement between only the autoEF and autoMAPSE in recordings deemed usable by evaluation of the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

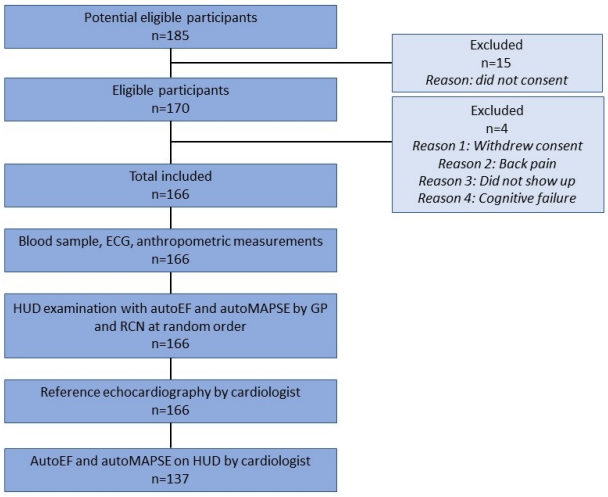

Figure 1 Study flow

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

338x190mm (96 x 96 DPI)

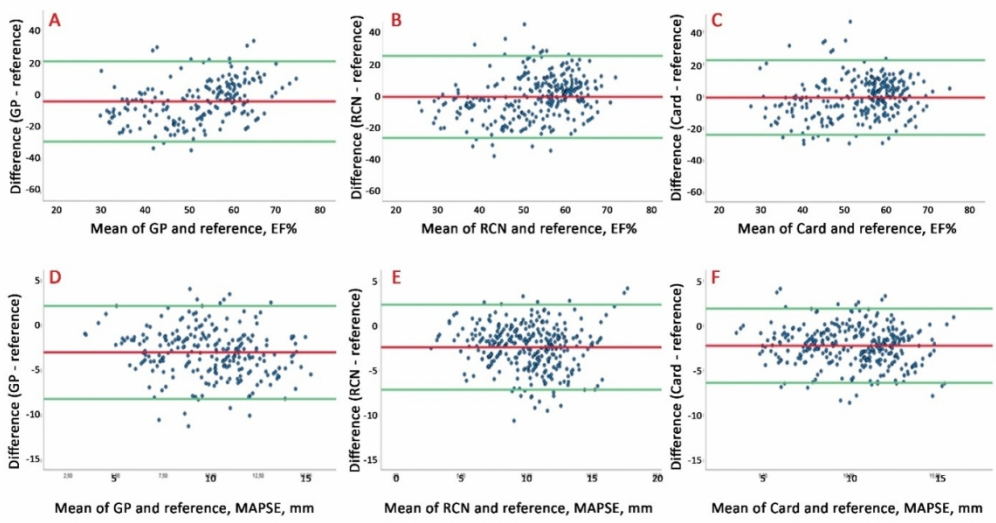

Figure 2. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in the total material.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography in the whole material (without exclusion of inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

160x84mm (220 x 220 DPI)

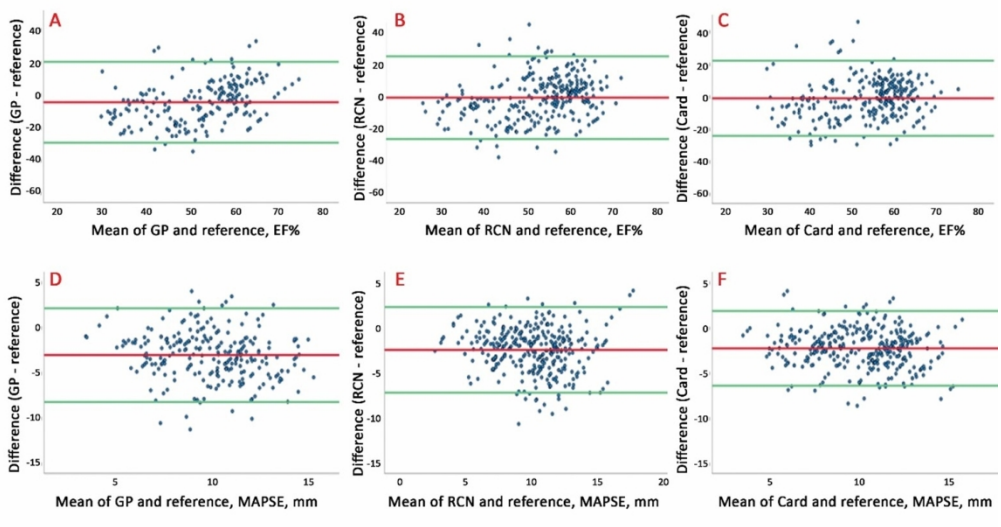

Figure 3. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in recordings not rejected by cardiologist.

Bland-Altman plots illustrating agreement between only the autoEF and autoMAPSE in recordings deemed usable by evaluation of the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

178x95mm (220 x 220 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	3-4
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	6
Participants	**6**	Eligibility criteria	6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	6
	9	Whether participants formed a consecutive, random or convenience series	6
Test methods	**10a**	Index test, in sufficient detail to allow replication	7-11
	10b	Reference standard, in sufficient detail to allow replication	7-11
	11	Rationale for choosing the reference standard (if alternatives exist)	7-11
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-11
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	7-11
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	7-11
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7-11
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	11
	15	How indeterminate index test or reference standard results were handled	11
	16	How missing data on the index test and reference standard were handled	11
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	11
	18	Intended sample size and how it was determined	11
RESULTS			
Participants	**19**	Flow of participants, using a diagram	12, 26
	20	Baseline demographic and clinical characteristics of participants	12, 22
	21a	Distribution of severity of disease in those with the target condition	12
	21b	Distribution of alternative diagnoses in those without the target condition	12
	22	Time interval and any clinical interventions between index test and reference standard	12
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	12-14
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	12-14
	25	Any adverse events from performing the index test or the reference standard	12-14
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	14-16
	27	Implications for practice, including the intended use and clinical role of the index test	14-16
OTHER INFORMATION			
	28	Registration number and name of registry	17
	29	Where the full study protocol can be accessed	Attatchment
	30	Sources of funding and other support; role of funders	17

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-063793.R1
Article Type:	Original research
Date Submitted by the Author:	23-Aug-2022
Complete List of Authors:	Hjorth-Hansen, Anna; Helse Nord-Trøndelag HF, Internal medicine; Norges teknisk-naturvitenskapelige universitet, MI Lab and Department of Circulation and Medical Imaging Magelssen, Malgorzata; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Andersen, Garrett; Levanger Hospital, Internal Medicine Graven, Torbjørn; Levanger Hospital, Internal Medicine Kleinau, Jens; Levanger Hospital, Internal Medicine Landstad, Bodil; Levanger Hospital, Department of Research Løvstakken, Lasse; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Skjetne, Kyrre; Levanger Hospital, Internal Medicine Mjølstad, Ole; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Dalen, Havard; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging
Primary Subject Heading:	Cardiovascular medicine
Secondary Subject Heading:	Radiology and imaging
Keywords:	Heart failure < CARDIOLOGY, Echocardiography < CARDIOLOGY, Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS, Cardiovascular imaging < RADIOLOGY & IMAGING, Ultrasound < RADIOLOGY & IMAGING

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Hjorth-Hansen Anna Katarina M.D[1,2]

Magelssen Malgorzata Izabela M.D*[1,3]

Andersen Garrett Newton M.D, Ph.D.[2]

Graven Torbjørn M.D[2]

Kleinau Jens Olaf M.D[2]

Landstad Bodil J, Ph.D.[4,5]

Løvstakken Lasse MSc, Ph.D.[1]

Skjetne Kyrre M.D[2]

Mjølstad Ole Christian M.D, Ph.D.[1,3]

Dalen Håvard M.D, Ph.D. [1,2,3]

[1]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

[2]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway.

[3]Clinic of Cardiology, St. Olavs University Hospital, Trondheim, Norway.

[4]Department of Health Sciences, Mid Sweden University, Sweden.

[5]Unit of Research, Education and Development, Östersund Hospital, Östersund, Sweden.

Address of correspondence

Anna K Hjorth-Hansen

Annakh@ntnu.no

Department of Circulation and Medical Imaging,

Norwegian University of Science and Technology,

Box 8905,

7491 Trondheim,

Norway

Abstract

Objectives: To evaluate the feasibility and reliability of hand-held ultrasound (HUD) examinations with real-time automatic decision-making software for ejection fraction (autoEF) and mitral annular plane excursion (autoMAPSE) by novices (general practitioners), intermediate- (registered cardiac nurses) and expert users (cardiologists), respectively, compared to reference echocardiography by cardiologists in an outpatient cohort with suspected heart failure (HF).

Design: Feasibility study of a diagnostic test

Setting and participants: 166 patients with suspected HF underwent HUD examinations by five novices and three intermediate-skilled users including automatic measurements, and five experts performed HUD examinations with automatic measurements. HUD results were compared to a reference echocardiography. A blinded cardiologist scored the HUD recordings with automatic measurements as 1) discard, 2) accept, but adjust the measurement, or 3) accept the measurement as it is.

Primary outcome measure: The feasibility of automatic decision-making software for quantification of left ventricular function.

Results: The different users were able to run autoEF and autoMAPSE in of all patients. The feasibility for obtaining accepted images (score of ≥2) with automatic measurements ranged 50-91%. The feasibility was lowest for novices and highest for experts for both autoEF and autoMAPSE (p≤0.001). Large coefficients of variation and wide coefficients of repeatability indicate moderate agreement. The corresponding intraclass correlations (ICC) were moderate to good (ICC 0.51-0.85) for intra-, and poor (ICC 0.35-0.51) for inter-rater analyses.

The modest to poor agreement and reliability were not explained by the experience and competence of the users only.

Conclusion: Novices, intermediate and expert users were able to record four-chamber views for automatic assessment of autoEF and autoMAPSE using hand-held ultrasound devices. The modest feasibility, agreement and reliability do not warrant implementation into clinical practice until further refinement and clinical evaluation.

Trial registration number: NCT03547076 (ClinicalTrial.gov)

Keywords: Heart failure, ejection fraction, mitral annular plane systolic excursion, agreement, reliability, diagnostic

Strengths and limitations:

- To our knowledge, no study has evaluated automatic real-time quantification of left ventricular function on hand-held ultrasound devices by inexperienced users. The three user groups had different levels of experience, ranging from no previous experience to expert level.
- The inexperienced operators were recruited by their role in the municipality and not based on motivation for attending the study.
- There is no gold standard for evaluation of left ventricular function and echocardiographic measurements by experienced cardiologists were used as reference.

- An error was detected in the first software version of the autoEF decision-support software, and this experience may have affected the results for the revised software as well.
- The study sample is expected to provide adequate power for analyses.

Introduction

Heart failure (HF) is a severe condition with poor prognosis and reduced quality of life which constitutes a burden on the health care system with high costs (1) and 26 million patients affected worldwide (2). Echocardiography is the cornerstone imaging modality for diagnostics and patient follow-up. Correct diagnosing can be challenging. It is shown that (in-training) cardiology fellows inaccurately interpret echocardiograms (3) and delay in diagnosis may be present in up to 40% (4).

Estimation of left ventricular (LV) ejection fraction (EF) is required for classification and treatment of HF (5). Another robust and easily obtainable measure of LV function is mitral annular plane systolic excursion (MAPSE) which is quite sensitive for detection of LV dysfunction (6-8), even when EF is preserved. Semi-automatic quantification of LV EF has been available for some time, but automatic quantification of MAPSE is scarcely available (7).

Hand-held ultrasound devices (HUD) have been widely implemented in the medical field over the last decade and increasingly used by non-experts (9). So far, quantification of LV size and function by HUDs relies on visual evaluation only (10). Several studies have shown high feasibility and reliability for inexperienced users performing simple tasks by HUDs (11-15). The experience and skill of the operator is essential for more advanced measures such as assessment of LV function (15,16). Automatic measurement of LV EF (autoEF) from apical HUD recordings are now commercially available, and a novel method for real-time automatic measurement of MAPSE (autoMAPSE) is available on the GE Vscan Extend (GE Utrasound AS, Horten, Norway) for research purposes. This allow for real-time quantification of LV function by HUDs, and thus, it is a need to evaluate the feasibility and reliability in clinical scenarios by different users before implementing into clinical practise.

We aimed to evaluate the feasibility and reliability of HUD including real-time autoEF and autoMAPSE performed by user of different levels of experience in an outpatient cohort with suspected HF. Specifically, the novice, intermediate, and expert groups were represented by general practitioners (GPs), registered cardiac nurses (RCNs), and experienced cardiologists, respectively. Comprehensive echocardiography by experienced cardiologists served as reference.

Methods

Study design

Figure 1 indicates the flow of the study participants. The patients were examined by one of five (GPs) and by one of three RCNs at random order. GPs and RCNs, representing novice and intermediate level examiners, were blinded to each other's results. Reference echocardiography was performed by one of five cardiologists blinded to preceding examinations. An additional HUD examination was performed by the cardiologists (expert group). Due to logistic reasons, the first 29 patients were not examined by HUD by the cardiologist. No further follow-up or ultrasound examinations of the participants were performed during the study. The study was approved by the regional committee for medical and health research ethics (REK 2017/2054) and registered in the ClinicalTrial.gov database (NCT03547076)

Participants

Patients referred to Levanger Hospital, Norway, with suspected HF were available for inclusion. Exclusion criteria were age <18 years, known HF and pre vious cardiac imaging within the last decade. Eligible patients were included from June 2018 to June 2020.

Inclusion was paused from March to June 2020 due to the COVID-19 pandemic. All participants gave their informed, written consent prior to inclusion. Eligible patients were consecutively included. The study was performed in conformity with the Declaration of Helsinki.

Training and education of personnel

The conductors of the study had no influence on the selection of GPs for the study which were selected by the municipality administration based on the GPs' position in the two municipalities of Levanger and Verdal.

A total of six GPs underwent training in focused cardiac ultrasound by HUD aligned to European recommendations (10). One dropped out due to change of occupation, and thus, five GPs participated in the study. All GPs underwent six in-hospital training days with one-to-one supervision by one of two residents experienced in focused cardiac ultrasound, in addition to two evening lectures provided by experts in diagnostic ultrasound and echocardiography. The GPs had the opportunity to use a personal HUD without supervision from the first day of training, but for no longer than three months prior to inclusion. None of them received additional training prior to study start. Upon direct request, no GP considered himself/herself underprepared to start inclusion. Only one of the six had performed focused ultrasound examinations prior to training (N=7), and thus, the group represents inexperienced users. They performed in total median (range) 46 (45-68) examinations prior to first inclusion, where median (range) 10 (9-20) were unsupervised and 36 (31-43) supervised, respectively.

Three RCNs with experience from a nurse-led outpatient HF clinic represent intermediate experienced users. They had experience in evaluation of pleural effusion, the inferior vena cava and clinical signs in HF patients. They had also participated in previous studies with limited ultrasound examinations of the heart (17). The RCNs had completed a total of median (range) 118 (74-221) limited echocardiographic examinations before patient inclusion, and thus, they did not undergo the same systematic training as the GPs. They were instructed in how to use the HUD and initialize autoEF and autoMAPSE approximately 4 weeks prior to inclusion.

Five cardiologists experienced in echocardiography (median 18 (6 to 43) years of experience) were only instructed in how to initialize the automatic tools on the HUDs and were not provided any additional training. All cardiologists were certified by the national authorities.

Test method

Each patient underwent three HUD examinations in addition to the reference imaging. All HUD examinations were performed by a Vscan Extend with a sector probe, and similarly, reference echocardiography by a Vivid E9 or E95 scanner (GE Ultrasound AS, Horten, Norway) with a 1.4-4.6 MHz phased array transducer. All examinations were performed according to standard operating procedures and included 4-chamber recordings of the LV. The protocol for the GPs included parasternal long- and short-axis views, apical 4-chamber view, subcostal 4-chamber view, evaluation of the inferior vena cava and the pleural cavities. The recording of the inferior caval vein included both maximum and minimum dimension during a normal breathing. Pleural cavities were assessed in in sitting position, and in case of pleural effusion craniocaudal images were recorded. RCNs recorded the same above-

mentioned views, as well as apical 2-chamber and apical long-axis views, right ventricular focused 4-chamber view, and atrial focused recordings. Additionally, RCNs recorded colour Doppler images of the mitral, aortic, and tricuspid valve not related to the objectives of the current study. Cardiologists recorded 4-chamber view only on the HUD, but the reference echocardiography was comprehensive (18).

For all HUD examinations live cine-loops of at least one cardiac cycle were recorded. The software for autoEF or autoMAPSE was launched by the operator and the automatic analysed recordings were subsequently stored on the HUD. This was repeated aiming for 6 separate recordings for automatic analyses by autoEF (3 recordings) and autoMAPSE (3 recordings). All recorded views and analyses were stored and transferred without delay to a Tricefy® cloud-based server (Trice Imaging Inc., CA, USA).

Reference echocardiographic examinations were performed according to recommendations (18) in a separate room immediately after the examinations by the GPs and RCNs. All measurements reflect the average of at least three (five in the case of arrhythmia) cardiac cycles. Central methodology follows: All measurements were performed using EchoPAC SW Only, version 202 and 203 (GE Ultrasound). The LV endocardial borders were traced in end-diastole and end-systole in 4-chamber and 2-chamber view. LV volumes (end-diastolic and end-systolic) and EF was calculated based on the traces using the biplane Simpson's method. MAPSE was measured as the longitudinal displacement of the mitral annular septal and lateral points in reconstructed motion mode.

Details of the automatic tools for quantification of LV function and image analyses

Before storing of the loop of the 4-chamber view, the specific application (autoEF or autoMAPSE) was launched on the HUD. The automatic measurements of LV volumes and EF was done by the commercially available LVivo® app (DiA Imaging Analysis, Be'er Sheva, Israel). The app provides fully automatic edge detection and tracing of the endocardial border in standard apical 4-chamber views throughout the cardiac cycle. LV volume was estimated at end-diastole and end-systole and EF was calculated from the volume estimates. MAPSE was estimated by an automated algorithm tracking the mitral annular septal and lateral points using a LV model. Technical details of the method are described in a previous paper (19). Shortly, a Real-time Contour Tracking Library (RCTL) was used to process and track the LV movement and images (GE, Vingmed, Norway) using a non-uniform rational B-spline model (20). The mitral annular septal and lateral points of the model were returned from RCTL. The array of points was evaluated to locate the maximum mitral annular plane displacement. MAPSE was calculated at the septal and lateral mitral annular points and as averaged values. For both autoEF and autoMAPSE the 4-chamber view recording with the overlay of the results from the automatic algorithm was stored as described above.

All HUD recordings were made available for blinded analyses by external cardiologists experienced in echocardiography. These cardiologists scored all recordings with the automatic measurement overlay as one of the following categories: 1) Discard (not for clinical use), 2) Accept, but adjust the result according to suboptimal performance, or 3) Accept the result as it is. The scoring took both the quality of the recordings and the performance of the application used into account. Thus, if the recording was not representative for a 4-chamber view the score was lower. The latter part of the scoring was

based on identification and tracking of the endocardial border (autoEF), or mitral annular points (autoMAPSE) combined with the numerical output.

During the study we detected an error in the autoEF software, and thus, the LVivo app was revised by the vendor during the summer of 2019. In total 103 were analysed with the first version of the autoEF software and 63 patients with the revised software.

Other measurements

Blood samples were drawn the same day and analysed at the in-hospital accredited laboratory. Serum N-terminal pro-Brain Natriuretic Peptide (NT-pro-BNP), serum creatinine and estimated glomerular filtration rate (eGFR, calculated by the Cockcroft-Gault equation), as well as electrolyte status (serum sodium and serum potassium) and haemoglobin (g/dL) were measured. New York Heart Association (NYHA) functional classification was scored by the nurses, and body weight (kg), body height (cm) and blood pressure (mmHg) were measured. Anthropometric measurements were rounded up to the nearest multiple of 1.

Patient and public involvement

Patients were not involved in setting the research question or the outcome measures. However, patient user groups were involved in planning of the study period as well as the ways of informing the patients and the society of the study results.

Analyses

Continuous variables were expressed as mean and standard deviation (SD) or as median and interquartile range (IQR) as appropriate. Evaluation of normality was done by evaluation of

histograms and normality plots. Categorical variables are presented as frequencies and proportions. Student's *t*-test and Wilcoxon test were used for comparison of groups when appropriate, ANOVA with post-hoc Least Significant Difference (LSD) correction were used to compare the three user groups. A study was judged as feasible if both the following two critreria were present: First, the user was able to acquire data with the fully automatic decision-support software. Second, the cardiologists blinded score of the recordings with the automatic measurement overlay was at least 2 (indicating that the recording and automatic measurement was accepted for clinical use). Proportions were compared using the Chi square test and Fisher's exact test when appropriate. Reliability of the measurements was evaluated by intraclass correlations (ICC), where values <0.5 were considered poor, 0.5-0.75 moderate, 0.75-0.9 good and >0.9 excellent (19). The intra-rater reliability was calculated by a two-way mixed effect model defined by absolute agreement in the dataset of single measurements analysed by the automatic methods as repeated measurements from the same patient are assumed to be more similar to each other than measurements between patients (21). The inter-rater reliability was calculated with a two-way random model defined by absolute agreement in the dataset of average measurements analysed by both the GPs, nurses and cardiologists by HUDs compared to reference. The agreement with reference echocardiography was evaluated by coefficients of variation, coefficient of repeatability indicating the minimal detectable change and Bland-Altman statistics. A p-value <0.05 was considered statistically significant. Sample size was calculated based on estimates of diagnostic precision using Sample Power (SPSS, Inc., Chicago, IL, USA). A sample size of 104 was needed to detect a difference of <15% of correctly diagnosed patients with HF compared to reference. As the proportion of patients with HF was expected to be small, we adjusted to a sample size of 150. Due to the revision the AutoEF software the sample size

was further adjusted to 170 to account for the new software version. All statistical analyses were performed using IBM SPSS Statistics, version 27 (SPSS Inc, Chicago, IL, USA).

Results

Participants

Baseline characteristics shown in Table 1. In total, 185 patients were invited to participate, 170 were included and four (n=4) were excluded (no show (n=1), cognitive failure (n=1), withdrawal of consent (n=2)). Shortly, 166 participants were included (47% women), mean (interquartile range) age was 70 (63-78) years. NT-proBNP was above 125 ng/L in 101 (61%) with an overall mean of 705 ng/L. More than half the population (93 (55%)) was in NYHA class ≥II and had obesity or overweight (123 (74%)). Chronic pulmonary diseases were relatively rare (24 (15%)). Atrial fibrillation was known in 49 (29%) of the patients, and present at inclusion in 40 (23%).

Table 1. 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 (mmHg)	150 ± 22
Diastolic blood pressure (mmHg)	83 ± 11
Glomerular filtration rate (ml/min) *	67 ± 18 (15-125)
Haemoglobin (g/dL)	14.4 ± 1.5
N-terminal pro brain natriuretic peptide (ng/L)	705 ± 1219 (6-11309)

NYHA functional class	
I, n (%)	63 (37)
II, n (%)	80 (47)
III, n (%)	12 (7)
IV, n (%)	1 (1)
Diuretics, n (%)	41 (25)
Beta blockers, n (%)	51 (31)
Angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker, n (%)	32 (19)
Atrial fibrillation, n (%)	49 (29)
Chronic obstructive pulmonary disease/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 ± (interquartile range) and proportions are n (%). Medications listed refer to the current use.

Abbreviations: *Calculated by the Cockcroft-Gault equation.

Test results

Feasibility

The novices were able to record at least one 4-chamber image with autoEF and autoMAPSE in 134 (80%) and 153 (92%) patients, respectively. The corresponding numbers for the intermediate group were 151 (90%) and 161 (96%), respectively (difference versus novices, both p<0.001). The experts were able to obtain the same measurements by the HUD for

autoEF in 91% of the cases and autoMAPSE in 99% (difference versus the intermediate group, both p<0.001).

The proportion of images judged as feasible (score of ≥2) by the blinded cardiologist (Table 2) was lowest for novices, higher for the intermediate group and highest for experts for both autoEF and autoMAPSE (all p≤0.001). Overall, ≤53% of images with autoEF or autoMAPSE by novices were judged feasible, increasing to 84% and 85% for autoEF and autoMAPSE by experts, respectively. In analyses taking the two versions of the autoEF algorithm into account, the feasibility for autoEF improved after the update for all examiners ranging from 68% for novices to 91% for experts (Table 2). Only very few recordings with the automatic algorithm overlays were scored as 3: "Accept the result as it is". In total, n (%) for autoEF and autoMAPSE were 7 (2%) and 23 (5%) for novices, 13 (3%) and 52 (11%) for the intermediate group and 25 (7%) and 67 (17%) for experts. The proportion of recordings scored as 3 ("Result accepted as it is") using autoEF was lower using the revised autoEF algorithm in novices and experts, and in total only for all users after revision.

The time used for the focused cardiac ultrasound examination was mean (SD) 18 (7) min for novices and 23 (7) min for the intermediate group. The time used for the six recordings with the automatic measurements were mean (SD) 4 min 34 sec (2min 20 sec) for novices, 3 min 21 sec (1 min 52 sec) in the intermediate group and 2 min 21 sec (1 min 19 sec) in experts, respectively.

Table 2. Feasibility of image recording and the use of automatic applications.

	Hand-held ultrasound operator

	GP (novice)	RCN (intermediate)	Cardiologist (expert)
AutoEF, all patients	205/400 (51%)	296/442 (67%)	298/357 (84%)
AutoEF, first software version	100/246 (41%)	149/270 (55%)	148/193 (77%)
AutoEF, revised software version	105/154 (68%)	147/172 (85%)	150/164 (91%)
AutoMAPSE, all patients	248/471 (53%)	335/467 (72%)	333/391 (85%)

Data are presented as number of feasible/available recordings in total (%). Feasible recordings were defined as score of ≥2 (i.e., accepted with or without need for adjustments by the blinded cardiologist). Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; GP, general practitioner; RCN, registered cardiac nurse.

ReliabilityTable 3 shows the agreement of autoEF and autoMAPSE by the different users with reference. Shortly, the large coefficients of variability and large coefficients of repeatability for all three examiners indicate poor agreement between the automatic applications compared to reference. There was only a modest difference with respect to agreement between the operators. The minimal detectable change estimated from the coefficient of repeatability for autoEF and autoMAPSE ranged 24.2-21.5% and 5.0-4.1 mm, respectively. After the revision of the autoEF software, the minimal detectable change was somewhat improved but still approximately 20%.

Table 4 shows that intra-rater ICCs were moderate for all user groups with values <0.75 for all except for autoMAPSE by the intermediate group (0.85) and experts (ICC 0.83). The intra-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.51-

0.72. The intra-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.70-0.85, respectively.

The inter-rater ICC was poor (≤0.51) for both automatic applications and all users. Inter-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.43-0.51. The inter-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.35-0.51, respectively.

Figure 2 shows the Bland-Altman plots of the mean difference versus bias for all HUD recordings with automatic applications versus reference per operator. Similarly, Figure 3 is limited to images accepted (score 2 or 3) by the blinded cardiologist. Overall, the agreement was poor to moderate. We found no association of size of the measurement with agreement, but the limits of agreement were lower for the most experienced users (also shown in Table 3) and after excluding the images deemed too poor for clinical use (Figure 3).

Table 3. Mean values and the agreement of automatic hand-held ultrasound measurements of left ventricular function compared to reference.

	Hand-held ultrasound operator			
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**	**Reference echocardiography**
Mean and agreement, autoEF (all recordings)				
Mean (SD), %*	51.7 (10.1)	52.9 (9.6)	53.3 (9.5)	53.4 (10.1)
Coefficient of variation, %	15.4	13.3	12.0	-
Coefficient of repeatability, %*	24.0	24.2	21.5	-
Mean and agreement, autoEF (first software version, n=107)				

Mean (SD), %*	52.6 (11.6)	54.2 (10.3)	55.0 (10.4)	53.5 (10.0)
Coefficient of variation, %	14.8	13.5	11.2	-
Coefficient of repeatability, %*	24.7	24.6	21.4	-
Mean and agreement, autoEF (revised software version, n=63)				
Mean (SD), %*	50.8 (8.4)	51.0 (8.3)	51.6 (8.1)	54.7 (9.6)
Coefficient of variation, %	16.0	13.1	12.9	-
Coefficient of repeatability, %*	20.6	20.6	19.8	-
Mean and agreement, autoMAPSE (all patients)				
Mean of septal and lateral position				
Mean (SD), mm	9.8 (2.4)	10.1 (2.6)	10.2 (2.5)	11.4 (2.9)
Coefficient of variation, %	24.3	20.5	18.9	-
Coefficient of repeatability, mm	5.0	4.8	4.1	-

Comprehensive echocardiography by experienced cardiologists used as reference. *%-points. Abbreviations: As described in Table 2.

Table 4. Intra- and inter-rater reliability of automatic measurements of left ventricular function by HUDs according to operators.

	HUD measurements by		
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**
Intra-rater intraclass correlation (ICC)			
AutoEF	0.58*	0.51	0.72
AutoMAPSE	0.70*	0.85	0.83
Inter-rater intraclass correlation (ICC)			

AutoEF	0.44	0.43	0.51
AutoMAPSE	0.35	0.44	0.51

Intra-rater intraclass correlation (ICC) calculated from single recordings per patient with automatic quantification of left ventricular function. Inter-rater intraclass correlation based on average values per patient and operator. *ICC of two repeated measures as only 38 patients had only three repeated measures of autoEF and 50 of autoMAPSE, respectively. Abbreviations: As described in Table 2.

Discussion

This study is to our knowledge the first study to evaluate the feasibility and reliability of real-time automatic decision-support software for quantification of LV function by HUDs across novices, intermediate experienced users and experts. The main findings were: Firstly, that the feasibility of the applications was acceptable, and the experts had the highest feasibility. Secondly, the agreement with reference was poor to moderate, and even for the experts the agreement and reliability were barely within recommended ranges for clinical use.

Participants

The study population represents patients referred for cardiac examination to rule-in or rule-out HF in everyday practice. The novices underwent limited, but dedicated training. The intermediate group utilized focused cardiac ultrasound in in their clinical practice, and the experts were experienced with echocardiography and HUDs. The training of novices, as well as lack of additional training for the more advanced user groups, was in line with comparable

studies and present recommendations (10,22,23). Most of the patients were overweight or obese and comorbidities as atrial fibrillation and hypertension were common. Thus, both poor acoustics and atrial fibrillation (presented in 29%) could interfere with image acquisition and accuracy of the automatic measurements.

Feasibility

The ability to run the automatic decision-support software was high for autoEF and autoMAPSE with >80% and >92% success rate for performance by all user groups when no quality assessment of the recorded image or performance of the applications was performed. The proportions were lowest for the novices and highest for the experts. The feasibility of the autoEF application significantly improved after revision. However, after blinded quality assessment by the cardiologist the feasibility was markedly impaired for both applications. In novices 35-40% of the automatic decision-support software recordings were not recommended for clinical use. In the intermediate group and experts, the corresponding proportions were approximately 20% and 10%, respectively. Additionally, the presented feasibility was somewhat lower with the second version of the autoEF software, which may be caused by stricter rules for when the algorithm succeeded. Recently, automatic quantification of LV EF has been evaluated in a couple of studies by experienced users (15,24). One study evaluated the same autoEF software used by a cardiology resident after six months of echocardiographic training the automatic quantification of LV EF succeeded in 76 of 112 patients (68%)(24). The feasibility of the autoEF application significantly improved after revision for all user groups, indicating that this finding was unrelated to training effect. This also highlights the importance of comprehensive evaluation of diagnostic decision-

support software, also after revisions of the software and not only before introduction to the market. Even though the feasibility was significantly improved after revision of the autoEF algorithm, the proportion of recordings with the highest possible score in blinded review by the cardiologist was somewhat lower. The time consumption for the complete HUD examinations was mean 18-23 min for novices and the intermediate group, which we believe, is acceptable in the everyday practice in selected cases if the potential for clinical benefit is significant. However, the time use was significantly higher than in previous publications evaluating focused cardiac ultrasound by HUDs performed by more experienced users(11,15,25)

The intra- and inter-rater ICCs for novices and the intermediate group, were mainly lower than what would be recommended for clinical use (commonly used cut-off of 0.75)(26). For experts the ICCs were somewhat higher, but compared with reference only 0.51, and in intra-rater analyses 0.72-0.83, respectively. In a recent publication using another HUD platform by a single cardiologist for automatic quantification of LV EF the ICC was 0.91(15). Even though the presented data are not directly comparable, they may indicate that reliability was somewhat lower in the present study, even when the autoEF software was used by experienced cardiologists in the current study. Further, we find that image quality and operator experience alone cannot explain the only moderate intra-operator reliability among the experienced cardiologists. Future studies must address how the next generation automatic analyses of LV function will perform across users of varying level of experience.

The agreement was poor for automatic measurements of EF and MAPSE for all users. Even though the bias for autoEF was lower for the most experienced users, the agreement was poor to moderate for all subgroups. . In the recent publications by Filipiak-Strzecka and

Papadopoulou, the lower – upper limits of agreement with reference were -10 – 12 (EF %) and -16 – 13 (EF %), respectively. Thus, both studies found somewhat better agreement for LV EF compared to the presented limits of agreement shown in Figures 2 and 3, but neither the design nor the presented data are directly comparable.. For autoMAPSE the underestimation compared to reference was consistent and this finding replicates a previous study by our group(19). This highlights that the cut-off for pathology is not interchangeable between different methods. Suboptimal image acquisitions by inexperienced users partially explains the results, but importantly the agreement was suboptimal also in experts. This indicates that the algorithms behind the applications need refinement before incorporation as a reliable tool in everyday clinical work independently of the skills of the users.

The patients' perspective

For the patients' perspective it is important to provide correct diagnosis, and thus, treatment as soon as possible. This may reduce the suffer and improve quality of care. Moving advanced diagnostics to the patients' point-of-care may shorten time to diagnosis and improve care. As indicated by this study it is of outmost importance to thoroughly evaluate novel methodology before implementing changes into clinical practice, since findings further diagnostic work-up may be delayed in case of false negative findings.

Strengths and limitations

The main strengths of this study are the design study blinded examinations of the consecutive patients by three different user groups ranging from trained novices to experts,

blinded review of the feasibility of the automatic algorithms' performance, and the use of similar HUDs equipped with two relevant automatic decision-support software. The real-time automatic quantification of LV function on HUDs by inexperienced user with real-time feedback is to our knowledge not done before. Further, the novices were recruited by the municipality based on their role at various health care institutions and not on personal motivation to attend the study. This improves the generalisability but may have impaired the performance of the novices compared to the more experienced user groups. The adequately powered study is another strength.

The most important limitations relate to that no gold standard for evaluation of LV function exists. Thus, measurements of LV function by HUDs were compared to the expert comprehensive echocardiographic measurement. However, the distribution of feasibility and reliability across groups are less influenced by the lack of gold standard and we believe the blinded evaluation of all recordings with the automatic decision-support overlay provides valuable insight into the performance of the HUD and the automatic decision-support software across user groups. Another limitation which may have influences the performance of the autoEF software is the fact that the first version had an internal error detected during blinded image analyses. The reduced performance of the first version may in special have challenged the less experienced users and may also be of importance after software revision. However, the performance of the revised software even among experts indicates that the automatic decision-support software needs further refinement before broad clinical implementation.

Conclusion

Novice GPs , intermediate experienced RCNs and expert cardiologists were able to perform automatic analyses of left ventricular function by automatic applications implemented in hand-held ultrasound devices. However, these automatic applications showed poor to moderate agreement with reference and a modest reliability. While this study is a step in the right direction using novel technology to aid health-care providers in diagnostic decision-making, there is a need for more reliable methods before large-scale implementation into clinical practice.

Acknowledgements

The Department of Circulation and Medical Imaging, Norwegian University of Science and Technology hosts a research collaboration between university, hospitals, and various vendors funded by the Norwegian Research Council. GE Ultrasound is one of these partners but had no role in planning or performing this study. We want to thank clinicians and other employees at Nord-Trøndelag Hospital Trust for their support and for contributing to data collection in this research project.

Ethics approval statement

The study was approved by the regional committee for medical and health research ethics (REK 2017/2054) and registered in the ClinicalTrial.gov database (NCT03547076).

Author contributions

AKHH has contributed to protocol description, data collection, data analyses, manuscript draft and revision. MM has contributed to data collection, manuscript draft and revision. GA, TG, JOK, KS, and OCM have contributed to data acquisition and manuscript revision. BL has contributed to manuscript revision. LL has provided software development, contribution to study design and manuscript revision. HD is the main deveopier of study design, has provided data acquisition, and manuscript revision.

Competing interests

GE Ultrasound provided the HUD devices for loan through a research contract with the project leader (HD), but GE had no role in performance of the study. MIM, OCM, LL and HD hold positions in Centre for Innovative Ultrasound Solutions where GE Ultrasound is one of the industrial partners. LL acts as part-time consultant for GE Ultrasound.

Funding statement

The study was funded by grants from the European Interreg A initiatives (Norwegian-Swedish initiative), Research Council of Norway (Norges Forskningsråd) and Nord-Trøndelag Hospital Trust.

Data sharing statement

Data will be shared upon reasonable request.

References

1. Maggioni AP, Dahlstrom U, Filippatos G et al. EURObservational Research Programme: the Heart Failure Pilot Survey (ESC-HF Pilot). Eur J Heart Fail 2010;12:1076-84.
2. Ponikowski P, Anker SD, AlHabib KF et al. Heart failure: preventing disease and death worldwide. ESC Heart Fail 2014;1:4-25.
3. Zhitny V, Iftekhar N, Alexander L, Ahsan C. Cardiology Fellow Diagnostic Accuracy and Data Interpretation Outcomes: A Review of the Current Literature. Vascular Health and Risk Management 2020;Volume 16:429-435.
4. Bottle A, Kim D, Aylin P, Cowie MR, Majeed A, Hayhoe B. Routes to diagnosis of heart failure: observational study using linked data in England. Heart 2018;104:600-605.
5. Ponikowski P, Voors AA, Anker SD et al. 2016 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;37:2129-2200.
6. Willenheimer R, Cline C, Israelsson B. Left ventricular atrioventricular plane displacement: an echocardiographic technique for rapid assessment of prognosis in heart failure. Heart 1997:230-236.
7. Snare SR, Mjolstad OC, Orderud F, Haugen BO, Torp H. Fast automatic measurement of mitral annulus excursion using a pocket-sized ultrasound system. Ultrasound Med Biol 2011;37:617-31.
8. Vinereanu D, Nicolaides E, Tweddel AC, Fraser AG. "Pure" diastolic dysfunction is associated with long-axis systolic dysfunction. Implications for the diagnosis and classification of heart failure. Eur J Heart Fail 2005;7:820-8.
9. Díaz-Gómez JL, Ingelfinger JR, Mayo PH, Koenig SJ. Point-of-Care Ultrasonography. New England Journal of Medicine 2021;385:1593-1602.
10. Cardim N, Dalen H, Voigt JU 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;20:245-252.
11. Andersen GN, Graven T, Skjetne K et al. Diagnostic influence of routine point-of-care pocket-size ultrasound examinations performed by medical residents. J Ultrasound Med 2015;34:627-36.
12. Andersen GN, Viset A, Mjolstad OC, Salvesen O, Dalen H, Haugen BO. Feasibility and accuracy of point-of-care pocket-size ultrasonography performed by medical students. Bmc Medical Education 2014;14.
13. 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.
14. Lipczyńska M, Szymański P, Klisiewicz A, Hoffman P. Hand-Carried Echocardiography in Heart Failure and Heart Failure Risk Population: A Community Based Prospective Study. Journal of the American Society of Echocardiography 2011;24:125-131.
15. Panoulas VF, Daigeler AL, Malaweera AS 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.
16. Mjolstad OC, Andersen GN, Dalen H et al. Feasibility and reliability of point-of-care pocket-size echocardiography performed by medical residents. Eur Heart J Cardiovasc Imaging 2013;14:1195-202.
17. 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. Journal of Ultrasound in Medicine 2020;39:2313-2323.
18. 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.
19. 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. Journal of Ultrasound in Medicine 2020;40:341-350.
20. Storve S, Grue JF, Samstad S, Dalen H, Haugen BO, Torp H. Realtime Automatic Assessment of Cardiac Function in Echocardiography. IEEE Trans Ultrason Ferroelectr Freq Control 2016;63:358-68.
21. Detry MA, Ma Y. Analyzing Repeated Measurements Using Mixed Models. JAMA 2016;315:407-408.
22. Spencer KT, Kimura BJ, Korcarz CE, Pellikka PA, Rahko PS, Siegel RJ. Focused cardiac ultrasound: recommendations from the American Society of Echocardiography. J Am Soc Echocardiogr 2013;26:567-81.
23. Via G, Hussain A, Wells M et al. International evidence-based recommendations for focused cardiac ultrasound. J Am Soc Echocardiogr 2014;27:683.e1-683.e33.
24. Filipiak-Strzecka D, Kasprzak JD, Wejner-Mik P, Szymczyk E, Wdowiak-Okrojek K, Lipiec P. Artificial Intelligence-Powered Measurement of Left Ventricular Ejection Fraction Using a Handheld Ultrasound Device. Ultrasound Med Biol 2021;47:1120-1125.
25. Andersen GN, Haugen BO, Graven T, Salvesen O, Mjolstad OC, Dalen H. Feasibility and reliability of point-of-care pocket-sized echocardiography. Eur J Echocardiogr 2011;12:665-70.
26. Kleijn SA, Aly MF, Terwee CB, van Rossum AC, Kamp O. Reliability of left ventricular volumes and function measurements using three-dimensional speckle tracking echocardiography. Eur Heart J Cardiovasc Imaging 2012;13:159-68.

Figure legends.

Figure 1.

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

Figure 2.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography in the whole material (without exclusion of inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

Figure 3.

Bland-Altman plots illustrating agreement between only the autoEF and autoMAPSE in recordings deemed usable by evaluation of the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

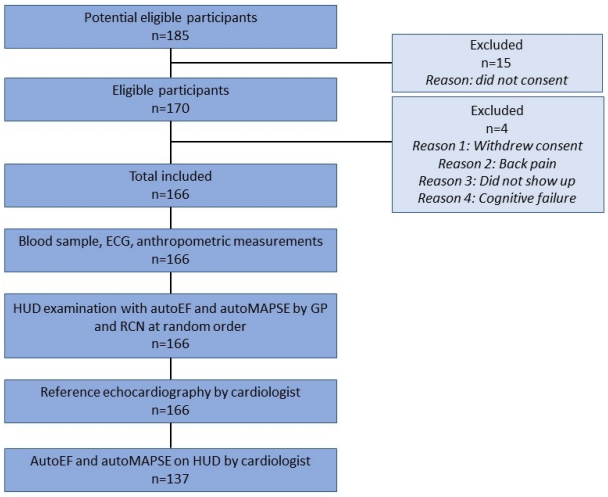

Figure 1 Study flow

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

338x190mm (96 x 96 DPI)

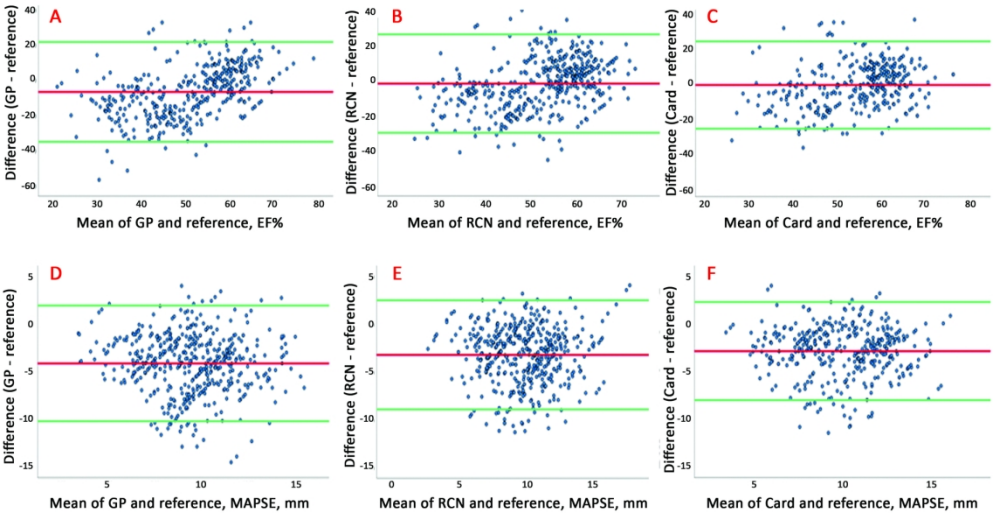

Figure 2. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in the total material.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography in the whole material (without exclusion of inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

161x85mm (300 x 300 DPI)

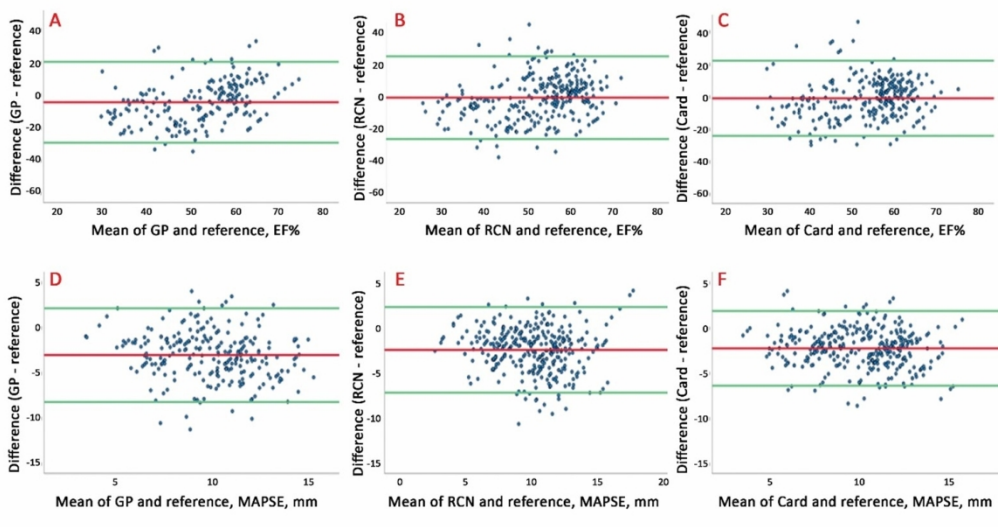

Figure 3. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in recordings not rejected by cardiologist.

Bland-Altman plots illustrating agreement between only the autoEF and autoMAPSE in recordings deemed usable by evaluation of the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

178x95mm (220 x 220 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	3-4
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	6
Participants	**6**	Eligibility criteria	6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	6
	9	Whether participants formed a consecutive, random or convenience series	6
Test methods	**10a**	Index test, in sufficient detail to allow replication	7-11
	10b	Reference standard, in sufficient detail to allow replication	7-11
	11	Rationale for choosing the reference standard (if alternatives exist)	7-11
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-11
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	7-11
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	7-11
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7-11
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	11
	15	How indeterminate index test or reference standard results were handled	11
	16	How missing data on the index test and reference standard were handled	11
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	11
	18	Intended sample size and how it was determined	11
RESULTS			
Participants	**19**	Flow of participants, using a diagram	12, 26
	20	Baseline demographic and clinical characteristics of participants	12, 22
	21a	Distribution of severity of disease in those with the target condition	12
	21b	Distribution of alternative diagnoses in those without the target condition	12
	22	Time interval and any clinical interventions between index test and reference standard	12
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	12-14
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	12-14
	25	Any adverse events from performing the index test or the reference standard	12-14
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	14-16
	27	Implications for practice, including the intended use and clinical role of the index test	14-16
OTHER INFORMATION			
	28	Registration number and name of registry	17
	29	Where the full study protocol can be accessed	Attatchment
	30	Sources of funding and other support; role of funders	17

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-063793.R2
Article Type:	Original research
Date Submitted by the Author:	29-Sep-2022
Complete List of Authors:	Hjorth-Hansen, Anna; Helse Nord-Trøndelag HF, Internal medicine; Norges teknisk-naturvitenskapelige universitet, MI Lab and Department of Circulation and Medical Imaging Magelssen, Malgorzata; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Andersen, Garrett; Levanger Hospital, Internal Medicine Graven, Torbjørn; Levanger Hospital, Internal Medicine Kleinau, Jens; Levanger Hospital, Internal Medicine Landstad, Bodil; Levanger Hospital, Department of Research Løvstakken, Lasse; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Skjetne, Kyrre; Levanger Hospital, Internal Medicine Mjølstad, Ole; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging Dalen, Havard; St Olavs Hospital Trondheim University Hospital, Cardiology; Norges teknisk-naturvitenskapelige universitet Fakultet for medisin og helsevitenskap, MI Lab and Department of Circulation and Medical Imaging
Primary Subject Heading:	Cardiovascular medicine
Secondary Subject Heading:	Radiology and imaging
Keywords:	Heart failure < CARDIOLOGY, Echocardiography < CARDIOLOGY, Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS, Cardiovascular imaging < RADIOLOGY & IMAGING, Ultrasound < RADIOLOGY & IMAGING

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Hjorth-Hansen Anna Katarina M.D[1,2]

Magelssen Malgorzata Izabela M.D*[1,3]

Andersen Garrett Newton M.D, Ph.D.[2]

Graven Torbjørn M.D[2]

Kleinau Jens Olaf M.D[2]

Landstad Bodil J, Ph.D.[4,5]

Løvstakken Lasse M.Sc., Ph.D.[1]

Skjetne Kyrre M.D[2]

Mjølstad Ole Christian M.D, Ph.D.[1,3]

Dalen Håvard M.D, Ph.D. [1,2,3]

[1]Department of Circulation and Medical Imaging (ISB), Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

[2]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway.

[3]Clinic of Cardiology, St. Olavs University Hospital, Trondheim, Norway.

[4]Department of Health Sciences, Mid Sweden University, Sweden.

[5]Unit of Research, Education and Development, Östersund Hospital, Östersund, Sweden.

Address of correspondence

Anna K Hjorth-Hansen

Annakh@ntnu.no

Department of Circulation and Medical Imaging,

Norwegian University of Science and Technology,

Box 8905,

7491 Trondheim,

Norway

Abstract

Objectives: To evaluate the feasibility and reliability of hand-held ultrasound (HUD) examinations with real-time automatic decision-making software for ejection fraction (autoEF) and mitral annular plane systolic excursion (autoMAPSE) by novices (general practitioners), intermediate- (registered cardiac nurses) and expert users (cardiologists), respectively, compared to reference echocardiography by cardiologists in an outpatient cohort with suspected heart failure (HF).

Design: Feasibility study of a diagnostic test

Setting and participants: 166 patients with suspected HF underwent HUD examinations with autoEF and autoMAPSE measurements by five novices, three intermediate-skilled users, and five experts. HUD results were compared to a reference echocardiography by experts. A blinded cardiologist scored all HUD recordings with automatic measurements as 1) discard, 2) accept, but adjust the measurement, or 3) accept the measurement as it is.

Primary outcome measure: The feasibility of automatic decision-making software for quantification of left ventricular function.

Results: The different users were able to run autoEF and autoMAPSE in all patients. The feasibility for obtaining accepted images (score of ≥2) with automatic measurements ranged 50-91%. The feasibility was lowest for novices and highest for experts for both autoEF and autoMAPSE ($p \leq 0.001$). Large coefficients of variation and wide coefficients of repeatability indicate moderate agreement. The corresponding intraclass correlations (ICC) were moderate to good (ICC 0.51-0.85) for intra-, and poor (ICC 0.35-0.51) for inter-rater analyses.

The findings of modest to poor agreement and reliability were not explained by the experience and competence of the users alone.

Conclusion: Novices, intermediate and expert users were able to record four-chamber views for automatic assessment of autoEF and autoMAPSE using hand-held ultrasound devices. The modest feasibility, agreement, and reliability suggest this should not be implemented into clinical practice without further refinement and clinical evaluation.

Trial registration number: NCT03547076 (ClinicalTrial.gov)

Keywords: Heart failure, ejection fraction, mitral annular plane systolic excursion, agreement, reliability, diagnostic

Strengths and limitations:

- To our knowledge, no study has evaluated automatic real-time quantification of left ventricular function on hand-held ultrasound devices by inexperienced users. The three user groups in this study had different levels of experience, ranging from no previous experience to expert level.
- The inexperienced operators were recruited by their role in the municipality and not based on motivation for attending the study.
- Due to the lack of a gold standard for evaluation of left ventricular function so echocardiographic measurements by experienced cardiologists were used as reference.
- An error detected in the first software version of the autoEF decision-support software may have affected the results for the revised software as well.

- The study sample is expected to provide adequate power for analyses.

Introduction

Heart failure (HF) is a severe condition with poor prognosis and reduced quality of life which constitutes a burden on the health care system with high costs and 26 million patients affected worldwide (1, 2). Echocardiography is the cornerstone imaging modality for HF diagnostics and patient follow-up. HF may be challenging to diagnose and it is shown that (in-training) cardiology fellows inaccurately interpret echocardiograms (3). Moreover, it is shown that a delayed HF diagnosis may be present in up to 40% of patients (4).

Estimation of left ventricular (LV) ejection fraction (EF) is required for classification and treatment of HF (5). Another robust and easily obtainable measure of LV function is mitral annular plane systolic excursion (MAPSE) which is quite sensitive for detection of LV dysfunction (6-8), even when EF is preserved. Semi-automatic quantification of LV EF has been available for some time, but automatic quantification of MAPSE is not widely available (7).

Hand-held ultrasound devices (HUD) have been widely implemented in the medical field over the last decade and are increasingly used by non-experts (9). So far, quantification of LV size and function by HUDs relies on visual evaluation only (10). Several studies have shown high feasibility and reliability for inexperienced users performing simple tasks by HUDs (11-15). The experience and skill of the operator is essential for more advanced measures such as assessment of LV function (15, 16). Automatic measurement of LV EF (autoEF) from apical HUD recordings are now commercially available, and a novel method for real-time automatic measurement of MAPSE (autoMAPSE) is available on the GE Vscan Extend (GE Ultrasound AS, Horten, Norway) for research purposes. This allows for real-time quantification of LV function by HUDs, and thus there is a need to evaluate the feasibility and reliability in clinical scenarios by different users before implementation into clinical practise.

We aimed to evaluate the feasibility and reliability of HUD examinations including real-time autoEF and autoMAPSE performed by users with different levels of experience in an outpatient cohort with suspected HF. Specifically, the novice, intermediate, and expert groups were represented by general practitioners (GPs), registered cardiac nurses (RCNs), and experienced cardiologists, respectively. Comprehensive echocardiography by experienced cardiologists served as reference.

Methods

Study design

Figure 1 indicates the flow of the study participants. The patients were examined by one of five GPs and by one of three RCNs at random order. GPs and RCNs were blinded to each other's results. Reference echocardiography was performed by one of five cardiologists blinded to preceding examinations. An additional HUD examination was performed by the cardiologists (expert group). Due to logistic reasons, the first 29 patients were not examined by HUD by the cardiologist. No additional follow-up or ultrasound examinations of the participants were performed related to the study. The study was approved by the regional committee for medical and health research ethics (REK 2017/2054) and registered in the ClinicalTrial.gov database (NCT03547076)

Participants

Patients referred to Levanger Hospital, Norway, with suspected HF were available for inclusion. Exclusion criteria were age <18 years, known HF and previous cardiac imaging within the last decade. Eligible patients were included from June 2018 to June 2020. Inclusion was paused from March to June 2020 due to the COVID-19 pandemic. All

participants gave their informed, written consent prior to inclusion. Eligible patients were consecutively included. The study was performed in conformity with the Declaration of Helsinki.

Training and education of personnel

The conductors of the study had no influence on the selection of GPs for the study who were selected by the municipality administration based on their position in the two municipalities of Levanger and Verdal.

A total of six GPs underwent training in focused cardiac ultrasound by HUDs aligned to the European recommendations (10). One dropped out due to change of occupation, and thus, five GPs participated in the study. All GPs underwent six in-hospital training days with one-to-one supervision by one of two residents experienced in focused cardiac ultrasound, in addition to two evening lectures provided by experts in diagnostic ultrasound and echocardiography. The GPs had the opportunity to use a personal HUD without supervision from the first day of training, but for no longer than three months prior to inclusion. None of them received additional training prior to study start. Upon direct request, no GP considered himself/herself underprepared to start inclusion. Only one of the six had performed focused ultrasound examinations prior to training (n=7 examinations), and thus, the group represents inexperienced users. They performed in total median (range) 46 (45-68) examinations prior to the first inclusion, where median (range) 10 (9-20) examinations were unsupervised and 36 (31-43) supervised, respectively.

Three RCNs with experience from a nurse-led outpatient HF clinic represented intermediate experienced users. They had experience in evaluation of pleural effusion, the inferior caval

vein, and evaluation of clinical signs in HF patients. Moreover, they had previously participated in studies with limited ultrasound examinations of the heart (17). The RCNs had completed a total of median (range) 118 (74-221) limited echocardiographic examinations before patient inclusion, and therefore, they did not undergo the same systematic training as the GPs. They were instructed on how to use the HUD and initialize the autoEF and autoMAPSE software approximately 4 weeks prior to inclusion.

Five cardiologists experienced in echocardiography (median 18 (6 to 43) years of experience) were only instructed in how to initialize the automatic decision-support software on the HUDs and were not provided any additional training. All cardiologists were certified by the national authorities.

Test method

Each patient underwent three HUD examinations in addition to the reference imaging. All HUD examinations were performed by a Vscan Extend with a sector probe, and similarly, reference echocardiography by a Vivid E9 or E95 scanner (GE Ultrasound AS, Horten, Norway) with a 1.4-4.6 MHz phased array transducer. All examinations were performed according to standard operating procedures and included 4-chamber recordings of the LV. The protocol for the GPs included parasternal long- and short-axis views, apical four-chamber view, subcostal four-chamber view, and evaluation of the inferior caval vein and the pleural cavities. The recording of the inferior caval vein included both maximum and minimum dimension during a normal breathing. Pleural cavities were assessed in sitting position, and in case of pleural effusion craniocaudal images were recorded. RCNs recorded the same above-mentioned views, as well as apical two-chamber and apical long-axis views,

right ventricular focused four-chamber view, and atrial focused recordings. Additionally, RCNs recorded colour Doppler images of the mitral, aortic, and tricuspid valve not related to the objectives of the current study. Cardiologists recorded the four-chamber view only by the HUD, but the reference echocardiography was comprehensive (18).

For all HUD examinations live cine-loops of at least one cardiac cycle were recorded. The software for autoEF or autoMAPSE implemented on the HUD was launched by the operator and the automatic analysed recordings were subsequently stored on the HUD. This was repeated aiming for six separate recordings for automatic analyses by autoEF (three recordings) and autoMAPSE (three recordings). All recorded views and analyses were stored and transferred without delay to a Tricefy® cloud-based server (Trice Imaging Inc., CA, USA). Reference echocardiographic examinations were performed according to recommendations (18) in a separate room immediately after the examinations by the GPs and RCNs. All measurements reflect the average of at least three (five in the case of arrhythmia) cardiac cycles. Central methodology follows: All measurements were performed using EchoPAC, version 202 and 203 (GE Ultrasound). The LV endocardial borders were traced in end-diastole and end-systole in 4-chamber and 2-chamber view. LV volumes (end-diastolic and end-systolic) and EF was calculated based on the traces using the biplane Simpson's method. MAPSE was measured as the longitudinal displacement of the mitral annular septal and lateral points in reconstructed motion mode.

Details of the automatic tools for quantification of LV function and image analyses

Before storing of the four-chamber view recording, the specific application (autoEF or autoMAPSE) was launched on the HUD. The automatic measurements of LV volumes and EF

were done by the commercially available LVivo® app (DiA Imaging Analysis, Be'er Sheva, Israel). The app provides fully automatic edge detection and tracing of the endocardial border in standard apical four-chamber views throughout the cardiac cycle. LV volume was estimated at end-diastole and end-systole and EF was calculated from the volume estimates. MAPSE was estimated by an automated algorithm tracking the mitral annular septal and lateral points using a LV model. Technical details of the method are described in a previous paper (19). Shortly, a Real-time Contour Tracking Library (RCTL) was used to process and track the LV movement and images (GE, Vingmed, Norway) using a non-uniform rational B-spline model (20). The mitral annular septal and lateral points of the model were returned from the RCTL. The array of points were evaluated to locate the maximum mitral annular plane displacement. MAPSE was calculated at the septal and lateral mitral annular points and as averaged values. For both autoEF and autoMAPSE the four-chamber view recording with the overlay of the results from the automatic algorithm was stored as described above.

All HUD recordings were made available for blinded analyses by external cardiologists experienced in echocardiography. These cardiologists scored all recordings with the automatic measurement overlay as one of the following categories: 1) Discard (not for clinical use), 2) Accept, but adjust the result according to suboptimal performance, or 3) Accept the result as it is. The scoring took both the quality of the recordings and the performance of the application used into account. Thus, if the recording was not representative for a four-chamber view the score was lower. The latter part of the scoring was based on identification and tracking of the endocardial border (autoEF), or mitral annular points (autoMAPSE) combined with the numerical output.

During the study we detected an error in the autoEF software, so the LVivo app was revised by the vendor during the summer of 2019. In total 103 were analysed with the first version of the autoEF software and 63 patients with the revised software.

Other measurements

Blood samples were drawn the same day and analysed at the in-hospital accredited laboratory. Serum N-terminal pro-Brain Natriuretic Peptide (NT-pro-BNP), serum creatinine and estimated glomerular filtration rate (eGFR, calculated by the Cockcroft-Gault equation), as well as serum electrolyte (sodium and potassium) and haemoglobin (g/dL) were measured. New York Heart Association (NYHA) functional classification was scored by the nurses, and body weight (kg), body height (cm) and blood pressure (mmHg) were measured. Anthropometric measurements were rounded up to the nearest multiple of one.

Patient and public involvement

Patients were not involved in decisions regarding the research question or the outcome measures. However, the patient user group was involved in planning of the study period as well as the ways of informing the patients and the society of the study results.

Analyses

Continuous variables were expressed as mean and standard deviation (SD) or as median interquartile range (IQR) as appropriate. Evaluation of normality was done by evaluation of histograms and normality plots. Categorical variables are presented as frequencies and proportions. Student's *t*-test and Wilcoxon test were used for comparison of groups when

appropriate, ANOVA with post-hoc Least Significant Difference (LSD) correction was used to compare the three user groups. A study was judged as feasible if the following two criteria were present: First, the user was able to acquire data with the fully automatic decision-support software. Second, the cardiologists blinded score of the recordings with the automatic measurement overlay was at least 2 (indicating that the recording and automatic measurement was accepted for clinical use). Proportions were compared using the Chi square test and Fisher's exact test when appropriate. Reliability of the measurements was evaluated by intraclass correlations (ICC), where values <0.5 were considered poor, 0.5-0.75 moderate, 0.75-0.9 good and >0.9 excellent (19). The intra-rater reliability was calculated by a two-way mixed effect model defined by absolute agreement in the dataset of single measurements analysed by the automatic methods as repeated measurements from the same patient are assumed to be more similar to each other than measurements between patients (21). The inter-rater reliability was calculated with a two-way random model defined by absolute agreement in the dataset of average measurements analysed by both the GPs, nurses and cardiologists by HUDs compared to reference. The agreement with reference echocardiography was evaluated by coefficients of variation, coefficient of repeatability indicating the minimal detectable change and Bland-Altman statistics. A p-value <0.05 was considered statistically significant. Sample size was calculated based on estimates of diagnostic precision using Sample Power (SPSS, Inc., Chicago, IL, USA). A sample size of 104 was needed to detect a difference of <15% of correctly diagnosed patients with HF compared to reference. As the proportion of patients with HF was expected to be small, we adjusted to a sample size of 150. Due to the revision the autoEF software the sample size was further adjusted to 170 to account for the new software version. All statistical analyses were performed using IBM SPSS Statistics, version 27 (SPSS Inc, Chicago, IL, USA).

Results

Participants

Baseline characteristics are shown in Table 1. In total, 185 patients were invited to participate, 170 were included and four (n=4) were excluded (no show (n=1), cognitive failure (n=1), withdrawal of consent (n=2)). The 166 participants included (47% women), median (interquartile range) age 70 (63-78) years. NT-proBNP was above 125 ng/L in 101 (61%) with an overall median of 295 ng/L. More than half the population was in NYHA class ≥II (93 (55%)) and were obese or overweight (123 (74%)). Chronic pulmonary diseases were relatively rare (24 (15%)). Atrial fibrillation was known in 49 (29%) of the patients, and present at inclusion in 40 (23%).

Table 1. 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 (mmHg)	150 ± 22
Diastolic blood pressure (mmHg)	83 ± 11
Glomerular filtration rate (ml/min) *	89 (68-109)
Haemoglobin (g/dL)	14.4 ± 1.5
N-terminal pro brain natriuretic peptide (ng/L)	295 (66-864)
NYHA functional class	
I, n (%)	63 (37)
II, n (%)	80 (47)

III, n (%)	12 (7)
IV, n (%)	1 (1)
Diuretics, n (%)	41 (25)
Beta blockers, n (%)	51 (31)
Angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker, n (%)	32 (19)
Atrial fibrillation, n (%)	49 (29)
Chronic obstructive pulmonary disease/asthma, n (%)	26 (16)
Diabetes mellitus type 2, n (%)	23 (14)
Coronary artery disease, n (%)	19 (11)

Normally distributed data are expressed as mean ± SD. Skewed data are presented as median ± (interquartile range). Proportions are presented as n (%). Medications refer to the current use. Abbreviations: *Calculated by the Cockcroft-Gault equation.

Test results

Feasibility

The novices were able to record at least one four-chamber image with autoEF and autoMAPSE in 134 (80%) and 153 (92%) patients, respectively. The corresponding numbers for the intermediate group were 151 (90%) and 161 (96%), respectively (difference versus novices, both p<0.001). The experts were able to obtain the same views using the HUD for autoEF in 91% of the cases and autoMAPSE in 99% (difference versus the intermediate group, both p<0.001).

The proportion of images judged as feasible (score of ≥2) by the blinded cardiologist was lowest for novices, higher for the intermediate group and highest for experts for both autoEF and autoMAPSE (all p≤0.001, Table 2). Overall, ≤53% of images with autoEF or autoMAPSE by novices were judged as feasible, compared to 84% and 85% for autoEF and autoMAPSE by experts, respectively. In analyses taking the two versions of the autoEF algorithm into account, the feasibility for autoEF improved after the revision for all examiners ranging from 68% for novices to 91% for experts (Table 2). Only very few recordings with the automatic algorithm overlays were scored as 3: "Accept the result as it is". In total, the numbers (%) for autoEF and autoMAPSE were 7 (2%) and 23 (5%) for novices, 13 (3%) and 52 (11%) for the intermediate group and 25 (7%) and 67 (17%) for experts. The proportion of recordings scored as 3 ("Result accepted as it is") using autoEF was lower using the revised autoEF algorithm in novices and experts.

The time used for the focused cardiac ultrasound examination was mean (SD) 18 (7) min for novices and 23 (7) min for the intermediate group. The time used for the six recordings with the automatic measurements were mean (SD) 4 min 34 sec (2min 20 sec) for novices, 3 min 21 sec (1 min 52 sec) in the intermediate group and 2 min 21 sec (1 min 19 sec) for experts, respectively.

Table 2. Feasibility (i.e., score ≥2) for the combinations of image recording and the use of automatic applications.

	Hand-held ultrasound operator		
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**
AutoEF, all patients	205/400 (51%)	296/442 (67%)	298/357 (84%)

AutoEF, first software version	100/246 (41%)	149/270 (55%)	148/193 (77%)
AutoEF, revised software version	105/154 (68%)	147/172 (85%)	150/164 (91%)
AutoMAPSE, all patients	248/471 (53%)	335/467 (72%)	333/391 (85%)

Data are presented as number of feasible/available recordings (%). Feasible recordings were defined as score of ≥2 (i.e., accepted with or without need for adjustments by the blinded cardiologist). Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; GP, general practitioner; RCN, registered cardiac nurse.

Reliability

Table 3 shows the agreement of autoEF and autoMAPSE by the different users with reference. In short, the large coefficients of variability and large coefficients of repeatability for all three user groups indicate poor agreement of the automatic applications compared to reference. There was only a modest difference with respect to agreement between the operators. The minimal detectable change estimated from the coefficient of repeatability for autoEF and autoMAPSE ranged 24.2-21.5%-points and 5.0-4.1 mm, respectively. After the revision of the autoEF software, the minimal detectable change was somewhat improved but still approximately 20%-points.

Table 4 shows that intra-rater ICCs were moderate for all user groups with values <0.75 for all except for autoMAPSE by the intermediate group (0.85) and experts (ICC 0.83). The intra-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.51-

0.72. The intra-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.70-0.85, respectively.

The inter-rater ICCs were poor (≤0.51) for both automatic decision support software and all users. Inter-rater ICC for autoEF was highest for experts, with ICCs for the three groups ranging 0.43-0.51. The inter-rater ICC for autoMAPSE was lowest for novices and highest for experts, with ICC ranging 0.35-0.51, respectively.

Figure 2 shows the Bland-Altman plots for HUD recordings with autoEF and autoMAPSE compared to reference according to user groups. Similarly, Figure 3 is limited to images accepted (score 2 or 3) by the blinded cardiologist. Overall, the agreement was poor to moderate. We found no association of size of the measurement with agreement, but the limits of agreement were lower for the most experienced users (also shown in Table 3) and after excluding the images deemed too poor for clinical use (Figure 3).

Table 3. Mean values and the agreement of automatic hand-held ultrasound measurements of left ventricular function compared to reference.

	Hand-held ultrasound operator			
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**	**Reference echocardiography**
Mean and agreement, autoEF (all recordings)				
Mean (SD), %*	51.7 (10.1)	52.9 (9.6)	53.3 (9.5)	53.4 (10.1)
Coefficient of variation, %	15.4	13.3	12.0	-
Coefficient of repeatability, %*	24.0	24.2	21.5	-
Mean and agreement, autoEF (first software version, n=107)				

Mean (SD), %*	52.6 (11.6)	54.2 (10.3)	55.0 (10.4)	53.5 (10.0)
Coefficient of variation, %	14.8	13.5	11.2	-
Coefficient of repeatability, %*	24.7	24.6	21.4	-
Mean and agreement, autoEF (revised software version, n=63)				
Mean (SD), %*	50.8 (8.4)	51.0 (8.3)	51.6 (8.1)	54.7 (9.6)
Coefficient of variation, %	16.0	13.1	12.9	-
Coefficient of repeatability, %*	20.6	20.6	19.8	-
Mean and agreement, autoMAPSE (all patients)				
Mean of septal and lateral position				
Mean (SD), mm	9.8 (2.4)	10.1 (2.6)	10.2 (2.5)	11.4 (2.9)
Coefficient of variation, %	24.3	20.5	18.9	-
Coefficient of repeatability, mm	5.0	4.8	4.1	-

Comprehensive echocardiography by experienced cardiologists used as reference. *%-points. Abbreviations: As described in Table 2.

Table 4. Intra- and inter-rater reliability of automatic measurements of left ventricular function by hand-held ultrasound according to operators.

	HUD measurements by		
	GP (novice)	**RCN (intermediate)**	**Cardiologist (expert)**
Intra-rater intraclass correlation (ICC)			
AutoEF	0.58*	0.51	0.72
AutoMAPSE	0.70*	0.85	0.83
Inter-rater intraclass correlation (ICC)			

AutoEF	0.44	0.43	0.51
AutoMAPSE	0.35	0.44	0.51

Intra-rater intraclass correlation (ICC) calculated from single recordings per patient with automatic quantification of left ventricular function. Inter-rater intraclass correlation based on average values per patient and operator. *ICC of two repeated measures as only few patients had three repeated measures of autoEF (n=38) and autoMAPSE (n=50), respectively. Abbreviations: As described in Table 2.

Discussion

This is to our knowledge the first study to evaluate the feasibility and reliability of real-time automatic decision-support software for quantification of LV function by HUDs across novices, intermediate experienced users, and experts. The main findings were: Firstly, that the feasibility of the applications was acceptable, even though being highest among experts. Secondly, the agreement with reference was poor to moderate, and even for the experts the agreement and reliability were barely within the ranges recommended for clinical use.

Participants

The study population represents patients referred for cardiac examination to rule-in or rule-out HF in everyday clinical practice. The novices underwent limited, but dedicated training. The intermediate group utilized focused cardiac ultrasound in in their clinical practice, and the experts were experienced in echocardiography and the use of HUDs. The training of novices, as well as lack of additional training for the more advanced user groups, was in line

with comparable studies and present recommendations (10, 22, 23). Most of the patients were overweight or obese and comorbidities such as atrial fibrillation and hypertension were common. Thus, both poor acoustics and atrial fibrillation (present during examination in 24%) could interfere with image acquisition and the precision of the automatic measurements.

Feasibility

The ability to run the automatic decision-support software was high for autoEF and autoMAPSE with >80% and >92% success rate for performance by all user groups when no quality assessment of the recorded image or performance of the applications was performed. The proportions were lowest for the novices and highest for the experts. The feasibility of the autoEF application significantly improved after revision. However, after blinded quality assessment by the external cardiologist the feasibility was markedly impaired for both applications. In novices 35-40% of the automatic decision-support software recordings were not recommended for clinical use. In the intermediate group and experts, the corresponding proportions were approximately 20% and 10%, respectively. Additionally, the proportion of images where the operators were able to run the autoEF software was somewhat lower with the second version of the software, which may be caused by stricter rules for when the algorithm succeeded. Recently, automatic quantification of LV EF has been evaluated in a couple of studies by experienced users (15, 24). One study evaluated the same autoEF software operated by a cardiology fellow trained in advanced echocardiography for six months prior to study start. There the automatic LV quantification succeeded in 76 of 112 patients (68%)(24). In our study the feasibility of the autoEF

application significantly improved after revision for all user groups. This finding indicates that the training effect was minimal. Our findings also highlight the importance of comprehensive evaluation of diagnostic decision-support software before implementation into clinical practise. This also applies to revised versions of the decision-support software and not only before introduction to the market. Additionally, the proportion of recordings with the highest possible score in blinded review by the cardiologist was somewhat lower after revision of the autoEF software. The time consumption for the complete HUD examinations was mean 18-23 min for novices and the intermediate group, which we believe is acceptable in the everyday practice in selected cases with significant potential for clinical benefit. However, the time use was higher than in previous publications evaluating focused cardiac ultrasound by HUDs performed by more experienced users (11, 15, 25)

The intra- and inter-rater ICCs for novices and the intermediate group were mainly lower than what would be recommended for clinical use (commonly used cut-off of 0.75)(26). For experts the ICCs were somewhat higher, but compared with reference only 0.51, and in intra-rater analyses 0.72-0.83, respectively. In a recent publication using another HUD platform by a single cardiologist for automatic quantification of LV EF the ICC was 0.91 (15). Even though the presented data are not directly comparable, they may indicate that reliability was somewhat lower in the present study, even when the autoEF software was used by experienced cardiologists in the current study. Further, we find that image quality and operator experience alone cannot fully explain the moderate intra-operator reliability among the experienced cardiologists. Future studies must address how the next generation automatic analyses of LV function will perform across users of varying level of experience.

The agreement was poor for automatic measurements of EF and MAPSE for all users. Even though the bias for autoEF was lower for the most experienced users, the agreement was poor to moderate for all user groups. In the recent publications by Filipiak-Strzecka and Papadopoulou, the lower – upper limits of agreement with reference were -10 – 12 (EF %) and -16 – 13 (EF %), respectively (24, 27). Thus, both studies found somewhat better agreement for LV EF compared to the presented limits of agreement shown in Figures 2 and 3, but neither the design nor the presented data are directly comparable. For autoMAPSE the underestimation compared to reference was consistent and replicates the findings from a previous study by our group (19). This highlights that the cut-off for pathology is not interchangeable between different methods. Suboptimal image acquisition by less experienced users partially explain the difference across user groups. Importantly the agreement and reliability were suboptimal also in experts which indicates that the decision-support software needs refinement before incorporation as a reliable tool in everyday clinical practice. The latter is of special importance before implementation by less experienced operators.

The patients' perspective

From the patients' perspective it is important to provide correct diagnosis, and thus, treatment as soon as possible. Fast and precise diagnostics may reduce patient suffering and improve the quality of care. Moving advanced diagnostics to the patients' point-of-care may shorten time to diagnosis and improve care. As indicated by this study it is of utmost importance to thoroughly evaluate novel methodology before implementation into clinical practice, since further diagnostic work-up may be delayed in case of false negative findings.

Strengths and limitations

The main strengths of this study design is the use of blinded examinations of the consecutive patients by three different user groups ranging from trained novices to experts, blinded review of the feasibility of the automatic algorithms' performance, and the use of similar HUDs equipped with two relevant automatic decision-support software. The real-time automatic quantification of LV function on HUDs by inexperienced users with real-time feedback has not to our knowledge been done before. Further, the novices were recruited by the municipality based on their role at various health care institutions and not on personal motivation to attend the study. This improves the generalisability but may have impaired the performance of the novices compared to the more experienced user groups. The adequate power of the study is another strength.

The most important limitation relates to the lack of a gold standard for evaluation of LV function. Thus, measurements of LV function by HUDs were compared to the experts' comprehensive echocardiographic measurements. However, the feasibility and reliability across groups are less influenced by the lack of a gold standard. Further, we believe that the blinded evaluation of all recordings with the automatic decision-support overlay provides valuable insight into the performance of the HUD and the automatic decision-support software across user groups. Another limitation which may have influenced the performance of the autoEF software is related to internal error of the first software version which was detected during blinded image analyses. The reduced performance of the first version may particularly have challenged the less experienced users and may also be of importance after software revision. However, the performance of the revised software (among experts)

indicates that the automatic decision-support software needs further refinement before broad clinical implementation.

Conclusion

Novice general practitioners, intermediate experienced registered cardiac nurses, and expert cardiologists were able to perform automatic analyses of left ventricular function by automatic decision-support software implemented in hand-held ultrasound devices. However, these automatic measurements showed poor to moderate agreement with reference and modest reliability. While this study is a step in the right direction using novel technology to aid health-care providers in diagnostic decision-making, there is a need for more reliable methods before large-scale implementation into clinical practice.

Acknowledgements

The Department of Circulation and Medical Imaging, Norwegian University of Science and Technology hosts a research collaboration between university, hospitals, and various vendors funded by the Norwegian Research Council. GE Ultrasound is one of these partners but had no role in planning or performing this study. We want to thank all clinicians and other employees involved at Nord-Trøndelag Hospital Trust for their support and for contributing to data collection in this research project.

Ethics approval statement

The study was approved by the regional committee for medical and health research ethics (REK 2017/2054) and registered in the ClinicalTrial.gov database (NCT03547076).

Author contributions

AKHH has contributed to protocol description, data collection, data analyses, manuscript draft and revision. MM has contributed to data collection, manuscript draft and revision. GA, TG, JOK, KS, and OCM have contributed to data acquisition and manuscript revision. BL has contributed to manuscript revision. LL has provided software development, contribution to study design and manuscript revision. HD is the main developer of study design, has provided data acquisition, and manuscript revision.

Competing interests

GE Ultrasound provided the HUD devices for loan through a research contract with the project leader (HD), but GE had no role in performance of the study. MIM, OCM, LL and HD hold positions at Centre for Innovative Ultrasound Solutions (CIUS) where GE Ultrasound is one of the industrial partners. LL acts as part-time consultant for GE Ultrasound.

Funding statement

The study was funded by grants from the European Interreg A initiatives (Norwegian-Swedish initiative), Research Council of Norway (Norges Forskningsråd) and Nord-Trøndelag Hospital Trust.

Data sharing statement

Data will be shared upon reasonable request.

References

1. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, et al. Heart failure: preventing disease and death worldwide. ESC Heart Fail. 2014;1(1):4-25.
2. Maggioni AP, Dahlstrom U, Filippatos G, Chioncel O, Leiro MC, Drozdz J, et al. EURObservational Research Programme: the Heart Failure Pilot Survey (ESC-HF Pilot). Eur J Heart Fail. 2010;12(10):1076-84.
3. Zhitny V, Iftekhar N, Alexander L, Ahsan C. Cardiology Fellow Diagnostic Accuracy and Data Interpretation Outcomes: A Review of the Current Literature . Vascular Health and Risk Management. 2020;Volume 16:429-35.
4. Bottle A, Kim D, Aylin P, Cowie MR, Majeed A, Hayhoe B. Routes to diagnosis of heart failure: observational study using linked data in England. Heart. 2018;104(7):600-5.
5. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 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;37(27):2129-200.
6. R W, C C, B I. Left ventricular atrioventricular plane displacement: an echocardiographic technique for rapid assessment of prognosis in heart failure. Heart. 1997(78):230-6.
7. Snare SR, Mjolstad OC, Orderud F, Haugen BO, Torp H. Fast automatic measurement of mitral annulus excursion using a pocket-sized ultrasound system. Ultrasound Med Biol. 2011;37(4):617-31.
8. Vinereanu D, Nicolaides E, Tweddel AC, Fraser AG. "Pure" diastolic dysfunction is associated with long-axis systolic dysfunction. Implications for the diagnosis and classification of heart failure. Eur J Heart Fail. 2005;7(5):820-8.
9. Díaz-Gómez JL, Ingelfinger JR, Mayo PH, Koenig SJ. Point-of-Care Ultrasonography. New England Journal of Medicine. 2021;385(17):1593-602.
10. Cardim N, Dalen H, Voigt JU, Ionescu A, Price S, Neskovic AN, 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(3):245-52.
11. Andersen GN, Graven T, Skjetne K, Mjolstad OC, Kleinau JO, Olsen O, et al. Diagnostic influence of routine point-of-care pocket-size ultrasound examinations performed by medical residents. J Ultrasound Med. 2015;34(4):627-36.
12. Andersen GN, Viset A, Mjolstad OC, Salvesen O, Dalen H, Haugen BO. Feasibility and accuracy of point-of-care pocket-size ultrasonography performed by medical students. Bmc Medical Education. 2014;14.
13. Gundersen GH, Norekval TM, Haug HH, Skjetne K, Kleinau JO, Graven T, 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(1):29-34.
14. Lipczyńska M, Szymański P, Klisiewicz A, Hoffman P. Hand-Carried Echocardiography in Heart Failure and Heart Failure Risk Population: A Community Based Prospective Study. Journal of the American Society of Echocardiography. 2011;24(2):125-31.
15. Panoulas VF, Daigeler AL, Malaweera AS, Lota AS, Baskaran D, Rahman S, 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(4):323-30.
16. Mjolstad OC, Andersen GN, Dalen H, Graven T, Skjetne K, Kleinau JO, et al. Feasibility and reliability of point-of-care pocket-size echocardiography performed by medical residents. Eur Heart J Cardiovasc Imaging. 2013;14(12):1195-202.
17. Hjorth-Hansen AK, Andersen GN, Graven T, Gundersen GH, Kleinau JO, Mjølstad OC, 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. Journal of Ultrasound in Medicine. 2020;39(12):2313-23.
18. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015;28(1):1-39 e14.
19. Magelssen MI, Palmer CL, Hjorth-Hansen A, Nilsen HO, Kiss G, Torp H, et al. Feasibility and Reliability of Automatic Quantitative Analyses of Mitral Annular Plane Systolic Excursion by Handheld Ultrasound Devices. Journal of Ultrasound in Medicine. 2020;40(2):341-50.
20. Storve S, Grue JF, Samstad S, Dalen H, Haugen BO, Torp H. Realtime Automatic Assessment of Cardiac Function in Echocardiography. IEEE Trans Ultrason Ferroelectr Freq Control. 2016;63(3):358-68.
21. Detry MA, Ma Y. Analyzing Repeated Measurements Using Mixed Models. JAMA. 2016;315(4):407-8.
22. Spencer KT, Kimura BJ, Korcarz CE, Pellikka PA, Rahko PS, Siegel RJ. Focused cardiac ultrasound: recommendations from the American Society of Echocardiography. J Am Soc Echocardiogr. 2013;26(6):567-81.
23. Via G, Hussain A, Wells M, Reardon R, ElBarbary M, Noble VE, et al. International evidence-based recommendations for focused cardiac ultrasound. J Am Soc Echocardiogr. 2014;27(7):683.e1-.e33.
24. Filipiak-Strzecka D, Kasprzak JD, Wejner-Mik P, Szymczyk E, Wdowiak-Okrojek K, Lipiec P. Artificial Intelligence-Powered Measurement of Left Ventricular Ejection Fraction Using a Handheld Ultrasound Device. Ultrasound Med Biol. 2021;47(4):1120-5.
25. Andersen GN, Haugen BO, Graven T, Salvesen O, Mjolstad OC, Dalen H. Feasibility and reliability of point-of-care pocket-sized echocardiography. Eur J Echocardiogr. 2011;12(9):665-70.
26. Kleijn SA, Aly MF, Terwee CB, van Rossum AC, Kamp O. Reliability of left ventricular volumes and function measurements using three-dimensional speckle tracking echocardiography. Eur Heart J Cardiovasc Imaging. 2012;13(2):159-68.
27. Papadopoulou S-L, Sachpekidis V, Kantartzi V, Styliadis I, Nihoyannopoulos P. 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. European Heart Journal - Digital Health. 2022.

Figure legends.

Figure 1.

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

Figure 2.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography on all recordings (without excluding inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

Figure 3.

Bland-Altman plots illustrating agreement between the autoEF and autoMAPSE only in recordings deemed usable for evaluation by the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

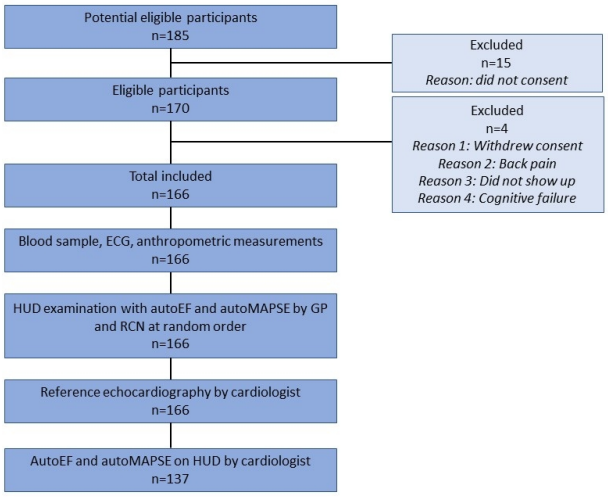

Figure 1 Study flow

Abbreviations: AutoEF, automatic measurement of left ventricular ejection fraction; AutoMAPSE, automatic measurement of mitral annular plane systolic excursion; ECG, electrocardiogram; HUD, hand-held ultrasound device; GP, general practitioner; RCN, registered cardiac nurse.

338x190mm (96 x 96 DPI)

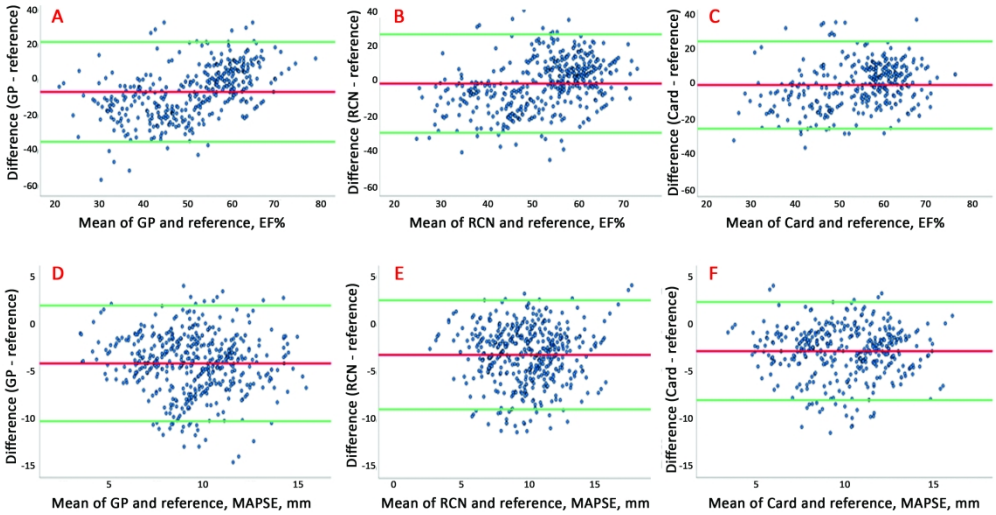

Figure 2. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in the total material.

Bland-Altman plots illustrating the agreement between all autoEF and autoMAPSE recordings taken by GPs, RCNs and cardiologists compared to reference echocardiography in the whole material (without exclusion of inacceptable recordings). Upper panel: autoEF by A) GPs, B) RCNs, and C) cardiologists compared to reference. Lower panel: autoMAPSE by D) GPs, E) RCNs, and F) cardiologists compared to reference. Abbreviations: Card, cardiologist; otherwise as in Figure 1.

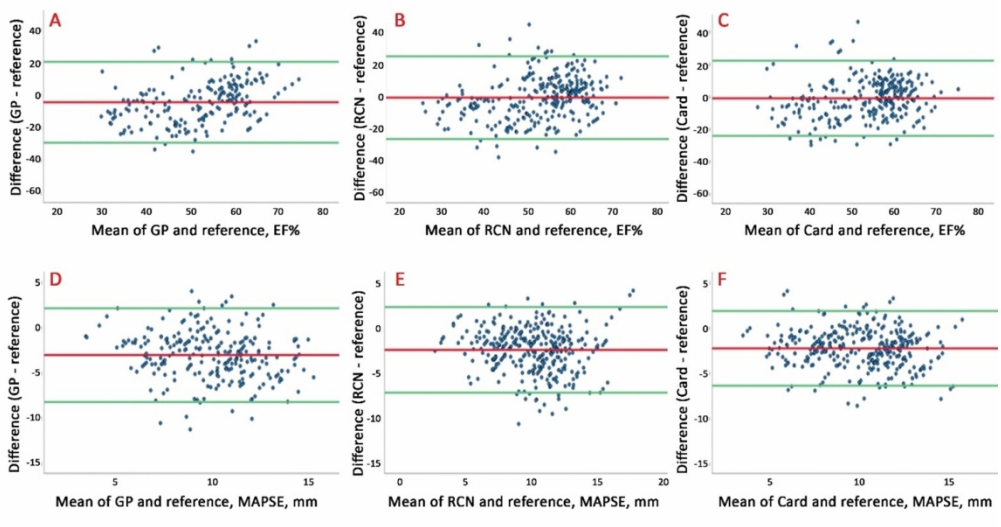

Figure 3. The agreement of automatic measurements of left ventricular function by automatic HUD applications compared to reference in recordings not rejected by cardiologist.

Bland-Altman plots illustrating agreement between only the autoEF and autoMAPSE in recordings deemed usable by evaluation of the blinded cardiologist. Upper panel; autoEF recorded by A) GPs, B) RCNs, and C) cardiologists. Lower panel; autoMAPSE by D) GPs, E) RCNs, and F) cardiologists. Abbreviations: As in Figure 1.

178x95mm (220 x 220 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	3-4
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	6
Participants	**6**	Eligibility criteria	6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	6
	9	Whether participants formed a consecutive, random or convenience series	6
Test methods	**10a**	Index test, in sufficient detail to allow replication	7-11
	10b	Reference standard, in sufficient detail to allow replication	7-11
	11	Rationale for choosing the reference standard (if alternatives exist)	7-11
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-11
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	7-11
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	7-11
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7-11
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	11
	15	How indeterminate index test or reference standard results were handled	11
	16	How missing data on the index test and reference standard were handled	11
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	11
	18	Intended sample size and how it was determined	11
RESULTS			
Participants	**19**	Flow of participants, using a diagram	12, 26
	20	Baseline demographic and clinical characteristics of participants	12, 22
	21a	Distribution of severity of disease in those with the target condition	12
	21b	Distribution of alternative diagnoses in those without the target condition	12
	22	Time interval and any clinical interventions between index test and reference standard	12
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	12-14
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	12-14
	25	Any adverse events from performing the index test or the reference standard	12-14
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	14-16
	27	Implications for practice, including the intended use and clinical role of the index test	14-16
OTHER INFORMATION			
	28	Registration number and name of registry	17
	29	Where the full study protocol can be accessed	Attatchment
	30	Sources of funding and other support; role of funders	17

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.