PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	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
AUTHORS	Hjorth-Hansen, Anna; Magelssen, Malgorzata; Andersen, Garrett; Graven, Torbjørn; Kleinau, Jens; Landstad, Bodil; Løvstakken, Lasse; Skjetne, Kyrre; Mjølstad, Ole; Dalen, Havard

VERSION 1 – REVIEW

REVIEWER	Arvig, Michael Slagelse Hospital, Emergency Medicine
REVIEW RETURNED	12-Jun-2022

GENERAL COMMENTS	Thanks for the opportunity to review this interesting and important article about a comparison of autoEF/MAPSE done by investigators with different levels of experience in ultrasound. As more and more automated ultrasound measurements are provided by the industry, it is essential to evaluate the diagnostic accuracy in a clinical context Comments and suggestions are arranged consecutively in the order the issues appear in the article. Title To be accurate, is this study not a feasibility study of a diagnostic test? Usually, in diagnostic accuracy studies, you calculate measurements like sensitivity, specificity, NPV, and PPV, which are not provided in this article. Reproducibility: Is this term used correctly? According to the COSMIN taxonomy, reproducibility consists of reliability, internal consistency, and measurement error. Is reliability a more correct nomenclature? Abstract Objective: Consider adding the disease and setting so the objective could stand alone.

	Abbreviations of GPs and RCNs are applied but are not used in the abstract. Consider either deleting it or using it in the subsequent paragraphs instead of novices and intermediate users. An overall feasibility score >80%: From just reading the abstract, I doubt what this exactly means. In the text, it seems that this number is just the number of times the investigators were able to acquire the A4C view and NOT obtain a score of 2 or more. Usually, the trial registration number in articles from BMJ Open is provided last in the abstract. Strengths and limitations American Society of Echocardiography level III: Maybe not all readers know what this level is. Is it the highest attainable? Maybe use another wording. Introduction Line 35, page 5: has=have. Line 52, page 5: GE Utrasound=Ultrasound. Line 54, page 5: allow=allows. Line 59, page 5. practise=practice. Aim: It should, in my opinion, could stand alone, so like in the abstract, I miss the setting and disease/patients. Again the wording of feasibility and reproducibility could be confusing. GPs and RCNs appear for the first time in the aim. So maybe already introduce the abbreviations here and not as provided in the method section. Methods Should the ethical approvement and registration on clinicaltrials.gov not be provided in the method section? Sample: Was it consecutive (besides the pause due to the pandemic) or randomly sampled from the referred patients with suspected HF? STARD item 9. The GPs underwent training in focused cardiac ultrasound, whereas the RCNs underwent training in limited echocardiography. Please, state if it was the same protocol. A lot of terminology regarding focused cardiac ultrasound exits. By focused cardiac ultrasound, do you mean the international guideline by Via? Then you could consider referencing it (Via G, Hussain A, Wells M, Reardon R, ElBarbary M, Noble VE, Tsung JW, Neskovic AN, Price S, Oren-Grinberg A, Liteplo A, Cordioli R, Naqvi N, Rola P, Poelaert J, Gulič TG, Sloth E, Labovitz A, Kimura B, Breitkreutz R, Masani N, Bowra J, Talmor D, Guarracinospan> F, Goudie A, Xiaoting W, Chawla R, Galderisi M, Blaivas M, Petrovic T, Storti E, Neri L, Melniker L; International Liaison Committee on Focused Cardiac UltraSound (ILC-FoCUS); International Conference on Focused Cardiac UltraSound (IC-FoCUS). International evidence-based recommendations for focused cardiac ultrasound. J Am Soc Echocardiogr. 2014 Jul;27(7):683.e1-683.e33. doi: 10.1016/j.echo.2014.05.001. PMID: 24951446).

	How were the cardiologists certified? And how experienced in years? Regarding inferior vena cava: Did the operators use the sniff test with a forced inspiration by the patient or just a “normal” breath cycle? Which probe was used, and what MHz? Why did the RCNS do a more extensive ultrasound with more views? The final measurement of EF was the autoEF and autoMAPSE, not more advanced EF calculations or eyeballing. Line 57, page 9. Technical details of the method is=are. How was the sample size calculated/determined? Line 24, page 11. Way=ways. Line 54, page 11. Was=were. Analyses: ANOVA with post-hoc LSD correction. Maybe a little explanation for this statistical test because it may be uncommon for some readers. Why do you use a mixed model for intrarater reliability? Of course, it is repeated measurements, but a short explanation could be helpful. And do you mean mixed effect model? And in Table 4, you write intrarater ICC? Results Excluded patients: Maybe add a “n” so the numbers are not confused with references, e.g. “no show (n=1)”. Age in mean? In my experience, age is seldom normally distributed. Table 1: It is stated under the table that “skewed data are presented as mean ±SD (interquartile range)”. But it must be median (IQR)? Line 36, page 14: overlay=overlays? Line 53, page 14: the automatic measurements was=were. Discussion Line 46, page 18. was=were. You state that the overall feasibility was >80% and >92%. However, these numbers are how many times the investigators were able to record at least one 4-chamber image with autoEF/MAPSE. But the feasibility is defined as a score of 2 or more? “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).” Please provide a reference for this statement. Consider making a separate subsection called, e.g., “Strengths and limitations”, to make the discussion more readable. Overall, in the discussion, I miss that this very interesting and important study is compared to the existing literature. Other studies: Are the results in line with other research? And if there are no previous studies, tell the reader (and maybe start your discussion by stating that “this is this first study to…”). How about the patient perspective of this study? Maybe fewer would be referred from GPs? Have you any number of how many of the referred patients with a suspicion of HF actually turn out to have HF? Conclusion: Why not use the abbreviations for GPs and RCNs already provided? So the conclusion could stand alone?

	References Refs 6 and 7 are a little bit old, maybe a newer source exits. 1

REVIEWER	Livesay, Georgia Princess Alexandra Hospital, Emergency
REVIEW RETURNED	10-Jul-2022

GENERAL COMMENTS	Study is well designed and carried out. Both abstract and limitations do not explore the extent to which the error in the software affects the results of the feasibility, which were poor for autoEF and worse with the revised algorithm. While you note that the poor agreement and reliability are not explained by experience and competence of users alone, this warrants further emphasis as a key finding: that the algorithms for the automated analyses are not yet sufficiently validated for use clinically and require further revision. Whatever the feasibility of achieving the images for the auto applications (and this was moderate at best for non-expert groups), if the software does not generate meaningful results the application is of no value. While the quality of written English is commendable, there are a small number of spelling, grammatical and syntax errors which require close editing if this paper is to be published in BMJOpen.

VERSION 1 – AUTHOR RESPONSE

Comments and replies – Reviewer 1

Title

1. To be accurate, is this study not a feasibility study of a diagnostic test? Usually, in diagnostic accuracy studies, you calculate measurements like sensitivity, specificity, NPV, and PPV, which are not provided in this article.

Authors reply: *See our response to the Editor's first comment. Changes are made in the revised abstract.*

2. Reproducibility: Is this term used correctly? According to the COSMIN taxonomy, reproducibility consists of reliability, internal consistency, and measurement error. Is reliability a more correct nomenclature?

Authors reply: *Reproducibility is commonly defined as the test results obtained by the several persons conducting the same test several times. In «A practical guide to assess the reproducibility of echocardiographic measures» (Bunting et al. J Am Soc Echocardiogr. 2019; 32:1505-15. doi: 10.1016/j.echo.2019.08.015) measures of agreement (Bland-Altman etc) and correlations/intraclass correlations relate to reproducibility, while coefficients of variation and minimal detectable change relate to reliability. In the COSMIN taxonomy the same three characteristics as described by the reviewer for reproducibility also relate to reliability (www.cosmin.nl/tools/cosmin-taxonomy-*

measurement-properties). As we performed repeated measurements on the same subject under very similar conditions, we have revised the terms used in the manuscript from reproducibility to reliability and added one sentence to explain the limitations in the discussion section.

Abstract

3. Objective: Consider adding the disease and setting so the objective could stand alone.

Authors reply: *We have revised the manuscript accordingly. Changes made in Abstract > Objectives.*

4. Abbreviations of GPs and RCNs are applied but are not used in the abstract. Consider either deleting it or using it in the subsequent paragraphs instead of novices and intermediate users.

Authors reply: *We have deleted the specified abbreviations in the abstract. Changes made in Abstract > Objectives.*

5. An overall feasibility score >80%: From just reading the abstract, I doubt what this exactly means. In the text, it seems that this number is just the number of times the investigators were able to acquire the A4C view and NOT obtain a score of 2 or more.

Authors reply: *We have rephrased these sentences to make the results clearer. Changes made in Abstract > Results.*

6. Usually, the trial registration number in articles from BMJ Open is provided last in the abstract.

Authors reply: *We have added the trial registration number to Abstract > Trial registration number.*

Strengths and limitations

7. American Society of Echocardiography level III: Maybe not all readers know what this level is. Is it the highest attainable? Maybe use another wording.

Authors reply: *We have rephrased the sentence to «expert level». Changes made in Strengths and limitations.*

Introduction

8. Line 35, page 5: has=have.

Authors reply: *The typo is corrected in revised manuscript.*

9. Line 52, page 5: GE Utrasound=Ultrasound.

Authors reply: *The typo is corrected in revised manuscript.*

10. Line 54, page 5: allow=allows.

Authors reply: *The typo is corrected in revised manuscript.*

11. Line 59, page 5. practise=practice.

Authors reply: *The typo is corrected in revised manuscript.*

12. Aim:
    a. It should, in my opinion, could stand alone, so like in the abstract, I miss the setting and disease/patients. Again, the wording of feasibility and reproducibility could be confusing.
    b. GPs and RCNs appear for the first time in the aim. So maybe already introduce the abbreviations here and not as provided in the method section.

Authors reply: *We have included the setting and the type of disease in the aim. Further, the abbreviations (GPs and RCNs) are introduced as first presented. Changes are made to the revised manuscript (Introduction; last paragraph).*

Methods

13. Should the ethical approvement and registration on clinicaltrials.gov not be provided in the method section?

Authors reply: *We have included details of the ethical approvement and clinical trial registration in the revised manuscript (Methods > Study design).*

14. Sample: Was it consecutive (besides the pause due to the pandemic) or randomly sampled from the referred patients with suspected HF? STARD item 9.

Authors reply: *Participants were consecutively included as no exclusion criteria were present. This information is added to the revised manuscript (Methods > Participants).*

15. The GPs underwent training in focused cardiac ultrasound, whereas the RCNs underwent training in limited echocardiography. Please, state if it was the same protocol. A lot of terminology regarding focused cardiac ultrasound exits. By focused cardiac ultrasound, do you mean the international guideline by Via? Then you could consider referencing it (Via G, Hussain A, Wells M, Reardon R, ElBarbary M, Noble VE, Tsung JW, Neskovic AN, Price S, Oren-Grinberg A, Liteplo A, Cordioli R, Naqvi N, Rola P, Poelaert J, Gulič TG, Sloth E, Labovitz A, Kimura B, Breitkreutz R, Masani N, Bowra J, Talmor D, Guarracino F, Goudie A, Xiaoting W, Chawla R, Galderisi M, Blaivas M, Petrovic T, Storti E, Neri L, Melniker L; International Liaison Committee on Focused Cardiac UltraSound (ILC-FoCUS); International Conference on Focused Cardiac UltraSound (IC-FoCUS). International evidence-based recommendations for focused cardiac ultrasound. J Am Soc Echocardiogr. 2014 Jul;27(7):683.e1-683.e33. doi: 10.1016/j.echo.2014.05.001. PMID: 24951446).

Authors reply: *The training of GPs were aligned to the position paper by the European Association of CardioVascular Imaging by Cardim N, Dalen H, Voigt JU, et al (https://doi.org/10.1093/ehjci/jey145) E-pub Oct 22, 2018. The project leader Dalen H was second author on the position paper, and thus, we were aware of the content of the paper during planning of the study. We have added this paper as reference for training of GPs and made minor revision to the description of training of the GPs. Changes made in the revised manuscript (Methods > Training and education of personnel).*

16. How were the cardiologists certified? And how experienced in years?

Authors reply: *All cardiologists were certified by the national (Norwegian) authority and had long experience (6-43 years, median 18 years) with echocardiography. This information is included in the revised manuscript (Methods > Training and education of personnel).*

17. Regarding inferior vena cava: Did the operators use the sniff test with a forced inspiration by the patient or just a “normal” breath cycle?

Authors reply: *The details of the inferior vena cava recordings are included in the revised manuscript (Methods > Test methods).*

18. Which probe was used, and what MHz?

Authors reply: *The details of the probes used are included in the revised manuscript (Methods > Test methods).*

19. Why did the RCNS do a more extensive ultrasound with more views? The final measurement of EF was the autoEF and autoMAPSE, not more advanced EF calculations or eyeballing.

Authors reply: *The RCNs had a more extensive protocol due to their previous experience with performing more comprehensive echocardiographic recordings in a previous study (reference 17). We considered qualitative evaluation of valvular pathology to be too extensive for the GPs to include in this study. Thus, the purpose of the additional collected data by the RCNs does not relate to the aims of this study, but to another sub study. As this was part of the data collection it is still included in the methods description. Minor changes are made in the revised manuscript (Methods > Test methods).*

20. Line 57, page 9. Technical details of the method is=are.

Authors reply: *The typo is corrected in revised manuscript.*

21. How was the sample size calculated/determined?

Authors reply: *Sample size was calculated using Sample Power (SPSS, Inc., Chicago, IL, USA). A difference of <15% of correctly diagnosed HF patients was considered to be of little importance. This resulted in a sample size of 104. As the number of patients with significant pathology was expected to be small, we originally planned for a sample size of 150 to account for a high number of normal findings. Throughout the study period, the autoEF software was upgraded due to an internal error and*

the sample size was adjusted to 170 to account for the new software version. This information is added to the revised manuscript. Changes made in Methods > Analyses.

22. Line 24, page 11. Way=ways.

Authors reply: *The typo is corrected in revised manuscript.*

23. Line 54, page 11. Was=were.

Authors reply: *The typo is corrected in revised manuscript.*

24. Analyses: ANOVA with post-hoc LSD correction. Maybe a little explanation for this statistical test because it may be uncommon for some readers.

Authors reply: *ANOVA determines if any of the three groups are different from each other. This means that we do not know which ones that differ. The LSD-correction allows to differentiate which ones that are different from each other. As explanations of the statistical methods used are rarely included in the statistical descriptions of this type of manuscripts, we have just made minor changes to the revised manuscript (Methods > Analyses).*

25. Why do you use a mixed model for intra-rater reliability? Of course, it is repeated measurements, but a short explanation could be helpful. And do you mean mixed effect model? And in Table 4, you write intrarater ICC?

Authors reply: *Repeated measurements from the same patient are assumed to be more similar to each other than measurements from one patient compared to another. Such correlation must be considered when analysing different measurements from the same patient. Mixed (effect) model for repeated measurements relating to missing data has the advantage of treating missing data in an unbiased manner and were chosen for this purpose. We have made minor changes to the revised manuscript (Methods > Analyses).*

Results

26. Excluded patients: Maybe add a "n" so the numbers are not confused with references, e.g. "no show (n=1)".

Authors reply: *We have revised the manuscript accordingly (Results > Participants).*

27. Age in mean? In my experience, age is seldom normally distributed.

Authors reply: *The age distribution is close to normally distributed in this population, however taking a closer look and reanalysing "age" we have chosen to take the mildly skewed distribution into consideration and report age as median (IQR). We have revised the manuscript accordingly (Results > Participants).*

28. Table 1: It is stated under the table that "skewed data are presented as mean ±SD (interquartile range)". But it must be median (IQR)?

Authors reply: *We have revised the logical error the manuscript accordingly (Results > Participants).*

29. Line 36, page 14: overlay=overlays?

Authors reply: *The typo is corrected in revised manuscript.*

30. Line 53, page 14: the automatic measurements was=were.

Authors reply: *The typo is corrected in revised manuscript.*

Discussion

31. Line 46, page 18. was=were.

Authors reply: *The typo is corrected in revised manuscript.*

32. You state that the overall feasibility was >80% and >92%. However, these numbers are how many times the investigators were able to record at least one 4-chamber image with autoEF/MAPSE. But the feasibility is defined as a score of 2 or more?

Authors reply: *We have rephrased the description of the proportion of exams able to run without no quality check-up to "the ability to run the automatic decision-support software" and specified the meaning of the term "feasibility" in the revised manuscript.*

33. "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)." Please provide a reference for this statement.

Authors reply: *We have added the following reference to the revised manuscript. Kleijn SA, Aly MFA, Terwee CB, et al. Reliability of left ventricular volumes and function measurements using three-dimensional speckle tracking echocardiography, Eur Heart J Cardiovasc Imaging. 2012; 13:159–168 PubMed . Doi: 10.1093/ejechocard/jer174*

34. Consider making a separate subsection called, e.g., "Strengths and limitations", to make the discussion more readable.

Authors reply: *We have added a subsection called* "Strengths and limitations" *to the revised manuscript.*

35. Overall, in the discussion, I miss that this very interesting and important study is compared to the existing literature.
    a. Other studies: Are the results in line with other research? And if there are no previous studies, tell the reader (and maybe start your discussion by stating that "this is this first study to…").

Authors reply: *We have revised the discussion section accordingly. See also our reply to the Editor's fourth comment.*

b. How about the patient perspective of this study? Maybe fewer would be referred from GPs? Have you any number of how many of the referred patients with a suspicion of HF actually turn out to have HF?

Authors reply: *The clinical importance of this study is a stand-alone publication currently under review. We have added a paragraph related to the patients' perspective at the end of the Discussion section.*

36. Conclusion: Why not use the abbreviations for GPs and RCNs already provided? So the conclusion could stand alone?

Authors reply: *We have revised the Conclusion accordingly.*

References

37. Refs 6 and 7 are a little bit old, maybe a newer source exists.

Authors reply: *Even though the included references are quite old, they are to our knowledge central manuscripts for the description of MAPSE. However, we have also added a recent and relevant reference in the revised manuscript. (Details of the added reference: Støylen A, Dalen H, Molmen HE. Left ventricular longitudinal shortening: relation to stroke volume and ejection fraction in ageing, blood pressure, body size and gender in the HUNT3 study. Open Heart. 2020 Sep;7(2):e001243. doi: 10.1136/openhrt-2020-001243.)*

Comments and replies – Reviewer 2

Study is well designed and carried out.

Both abstract and limitations do not explore the extent to which the error in the software affects the results of the feasibility, which were poor for autoEF and worse with the revised algorithm. While you note that the poor agreement and reliability are not explained by experience and competence of users alone, this warrants further emphasis as a key finding: that the algorithms for the automated analyses are not yet sufficiently validated for use clinically and require further revision.

Authors' reply: Unfortunately, we have been unclear describing the feasibility of the revised version. The feasibility (defined as score ≥2; indicating acceptable results in blinded review) was highly significant as shown in Results > Feasibility and Table 2. However, the proportion of recordings with the autoEF overlay scored as 3 (defined as "Accepted as it is", indicating perfect tracking and results of the algorithm) was somewhat lower for novices and experts. We have made changes to the Results > Feasibility and the Discussion > Feasibility sections in the revised manuscript.

Whatever the feasibility of achieving the images for the auto applications (and this was moderate at best for non-expert groups), if the software does not generate meaningful results the application is of no value.

Authors' reply: *We agree that the combination of a high feasibility, with sufficient agreement with reference and reliability is mandatory before implementing novel diagnostic tools into clinical practice.*

We believe the well-suited design, the wide range of operator experience, and repeated measurements all constitute strengths of the study and that it is important to inform the clinical and scientific society of this finding as HUDs are increasingly implemented in clinical practice and often used by inexperienced users. As shown by our results and the presented discussion we recommend that these supportive algorithms should not be used without further refinement and testing.
While the quality of written English is commendable, there are a small number of spelling, grammatical and syntax errors which require close editing if this paper is to be published in BMJOpen.
Authors' reply: *We have revised the manuscript accordingly and corrected all revealed typos.*

VERSION 2 – REVIEW

REVIEWER	Arvig, Michael Slagelse Hospital, Emergency Medicine
REVIEW RETURNED	26-Aug-2022

GENERAL COMMENTS	Dear Editor. Thank you for the opportunity to review a revised version of the manuscript. The manuscript has improved after revision, and all my raised points have been fully explained and addressed. I only have one very minor correction: 1. P 16, line 25: Now it states mean (IQR) = 70 (63-78). However, should it not be median (IQR) = 73 (63-78)? Best regards Michael Dan Arvig, MD

REVIEWER	Livesay, Georgia Princess Alexandra Hospital, Emergency
REVIEW RETURNED	21-Sep-2022

GENERAL COMMENTS	Reviewer comments for Manuscript ID: bmjopen-2022-063793.R1 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 Page numbers refer to those of the ScholarOne manuscript, not the paper itself Line numbers refer to those displayed on the manuscript, although these do not appear to line up well with the script. P5 line 48: in of all = in all

	P6 line 6: only = alone P6 line 14: do not warrant implementation = suggest this should not be implemented P6 line 17: until = without P6 line 49: and = so P8 line 14: diagnosing = diagnosis P8 line 32: scarcely = not widely P8 line 54: allow = allows P8 line 57: thus, it = thus there P8 line 59: implementing = implementation P9 line 28: (GPs) = GPs P10 line 18: which = who P11 line 3: represent = represented P17 line 56: measurements by = views using P18 title of Table 2 would be better as: Feasibility (ie feasibility score ≥ 2) of image recording for the use of automatic applications. P19 line 36: Shortly = In short P23 line 6: comorbidities as = comorbidities such as P25 line 37: For = from P25 line 39: the suffer = suffering P25 line 44: outmost = utmost

	P25 line 49: findings further = further P25 line 58: are the design study blinded examinations of the = design are the use of blinded examinations of P26 line 8: user = users P26 line 11: is to our knowledge not done = has not to our knowledge been done P26 line 23: to that no gold = to the lack of gold

VERSION 2 – AUTHOR RESPONSE

Comments and replies – Reviewer 1

Thank you for the opportunity to review a revised version of the manuscript. The manuscript has improved after revision, and all my raised points have been fully explained and addressed.

I only have one very minor correction:

1. P 16, line 25: Now it states mean (IQR) = 70 (63-78). However, should it not be median (IQR) = 73 (63-78)?

Authors' reply: *We have adjusted the description of age to median (IQR) throughout the revised manuscript.*

Comments and replies – Reviewer 2

Comments to the Author:
Your revisions have considerably improved this paper which now defines the objective and presents the results appropriately.
From my perspective no further major changes are required.
Several errors in the written English detract from the clarity of the paper and I have addressed these in the comments in the attached file.

Authors' reply: *We have corrected all typos included in the attachment. Some sentences are rephrased to improve the clarity of the content.*