UNIVERSITY OF WEST LONDON
The Career University

UWL REPOSITORY

repository.uwl.ac.uk

Utility of multi-expert approach for validation of AI in echocardiography

Chawla, S, Almeida, G, Ng, T S J, Stowell, C, Jevsikov, J, Manisty, C, Zolgharni, Massoud ORCID logoORCID: https://orcid.org/0000-0003-0904-2904, Shun-Shin, M and Francis, D P (2026) Utility of multi-expert approach for validation of AI in echocardiography. In: ESC Congress 2026, 28–31 August 2026, Munich, Germany.

http://dx.doi.org/10.1093/eurheartj/ehag785.007

This is the Published Version of the final output.

UWL repository link: https://repository.uwl.ac.uk/id/eprint/15731/

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Rights Retention Statement:

Utility of multi-expert approach for validation of AI in echocardiography

S. Chawla[1], G. Almeida[2], T.S.J. Ng[2], C. Stowell[2], J. Jevsikov[2], C. Manisty[3], M. Zolgharni[4], M. Shun-Shin[2], D.P. Francis[2]

[1]Imperial College Healthcare NHS Trust, London, United Kingdom of Great Britain & Northern Ireland
[2]Imperial College London, National Heart and Lung Institute, London, United Kingdom of Great Britain & Northern Ireland
[3]University College London, London, United Kingdom of Great Britain & Northern Ireland
[4]University of West London, London, United Kingdom of Great Britain & Northern Ireland

Funding Acknowledgements: Type of funding sources: Foundation. Main funding source(s): British Heart Foundation

Background: Historically, automatic measurement methods were validated by comparison to an expert human measurement. We study a multi-expert approach that may offer richer and clearer evaluation results.

Purpose: To evaluate multi-expert consensus as a reference to permit evaluation of AI performance that may be superhuman.

Methods: Six human experts, with mutual blinding, each directly measured 59 echocardiographic variables in 100 echocardiograms, and a further 113 variables were derived algebraically from these. An AI system also measured these variables.

For each variable, in each echocardiogram, we could therefore calculate how far Expert 1 was from the consensus of the other 5, and how far the AI was. We could then do the same for Expert 2, 3, 4, 5, and 6 in turn. This gave, for each variable, ~600 "expert deviations from consensus of 5 other experts", and ~600 "AI deviations from consensus of 5 human experts", which could then be compared by F-test.

We tested how informative this analysis would be with fewer experts, by reducing the dataset accordingly.

Finally, we tested a simple statistic that might allow unbiased evaluations even when some experts had missing data for some studies: comparison of distribution of pairwise differences in human-human versus AI-human pairings.

Results: When we reduced the dataset to only two human experts, the AI appeared to perform similarly to experts. For example, for LVEF, AI-human variance 26.3%², human-human variance 24.5%², ratio of variances = 1.07, p=.74 by F-test. However, when we increased the dataset to 3 experts the ratio of variances became 1.03, p=.88. With 4 experts, ratio = 0.86, p=.46. With 5 experts, ratio = 0.60, p=.01. With 6 experts, ratio = 0.46, p<.001 (Figure 1 left).

Of the 174 variables, with only 2 experts it was evident that the AI performed better than expert in 2/174 variables. With 3 experts, 18/174. With 4 experts, 25/174. With 5 experts, 32/174. With 6 experts, 40/174 (Figure 1 right).

Finally, the simple pairwise difference analysis measured performance without bias, even when 1%, 2%, 5%, 10%, 20% and 50% of the expert data are removed at random. Each run gave a slightly different ratio between AI-human and human-human differences, and this dispersion increased as data removal became more extensive (p<0.001), but the findings remained unbiased: centred on the true findings (full dataset). Figure 2 shows results for LVEF; other variables showed the same pattern.

Conclusion: AI echocardiographic measurements should be validated against multiple, mutually blinded, experts. This allows identification of AI methods that deliver better measurement performance than that delivered by any individual expert, which is increasingly possible with newer technologies. If a dataset has gaps, the pairwise difference comparison (between human-human and AI-human pairings) continues to provide a valid result and is easy to display and understand.

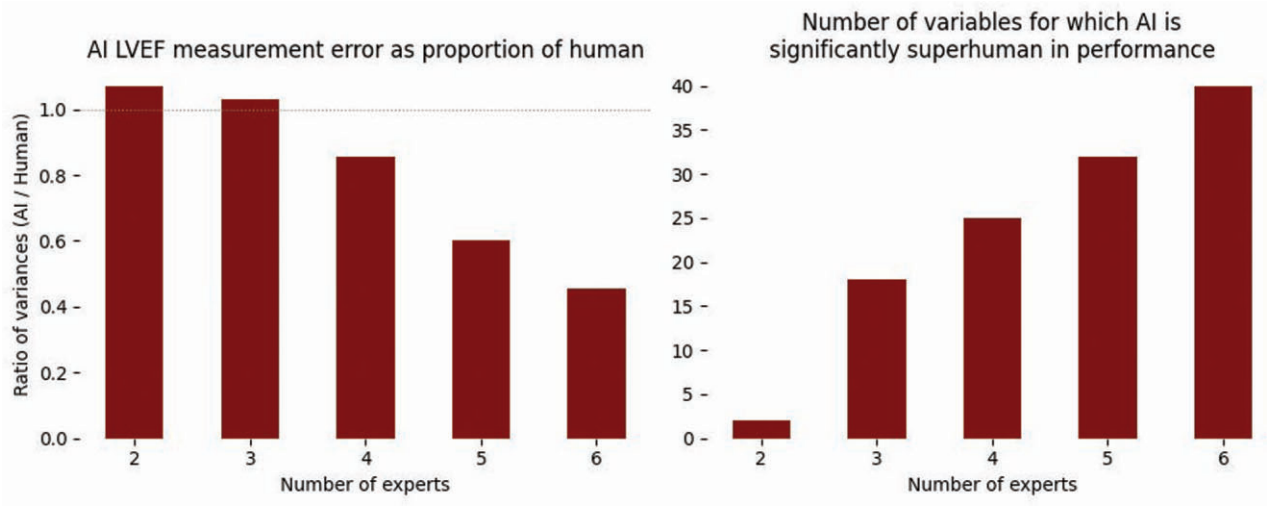


Figure 1

Downloaded from academic.oup.com/eurheartj/article/47/Supplement_1/ehag785.007/8997255 by guest on 31 October 2026

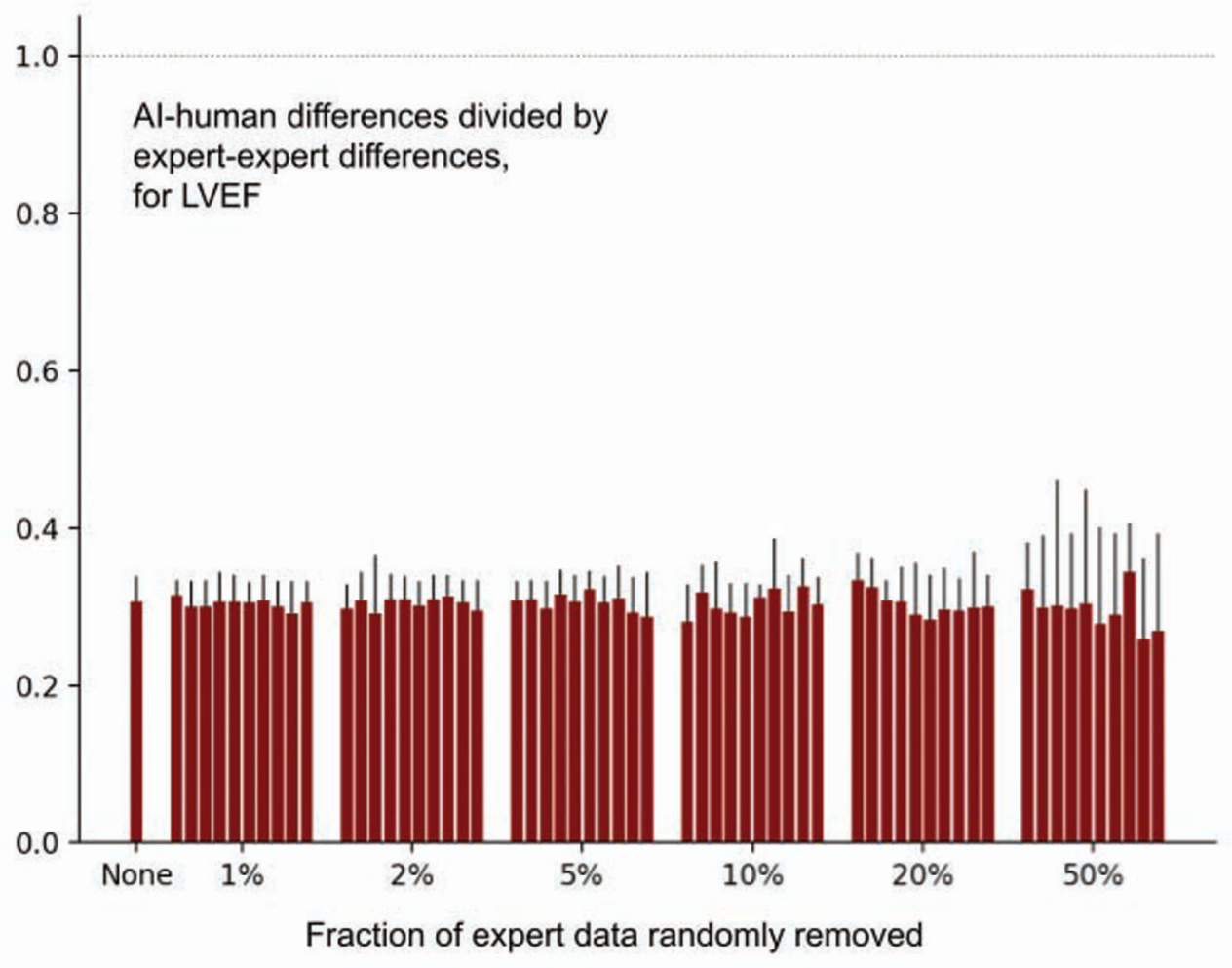


Figure 2

Downloaded from academic.oup.com/eurheartj/article/47/Supplement_1/ehag785.007/8997255 by guest on 31 October 2026