Independent and openly reported head-to-head comparative validation studies of AI medical devices: a necessary step towards safe and responsible clinical AI deployment

Lancet Digit Health 2025; 7: 100915

Published Online December 10, 2025
https://doi.org/10.1016/j.landig.2025.100915

See Articles https://doi.org/10.1016/j.landig.2025.100914

The increasing use of artificial intelligence (AI)-enabled medical devices has been accompanied by increasing evidence of a deficit between the performance claimed by the vendor and the performance experienced by users in the real world,[1,2] raising concerns regarding the safety and actual value of the devices. These concerns are particularly important in the context of national screening programmes procuring an AI system, which might affect entire populations and incur considerable cost with regard to procurement, integration with the existing information technology infrastructure, and staff training.

This problem was illustrated by a 2021 study by Lee and colleagues[3] in the USA. The authors evaluated the performance of seven different AI systems from five companies for automated diabetic retinopathy screening. This independent evaluation revealed two crucial issues. Firstly, the performance of these AI systems varied to a large extent, with sensitivity of detection of referable diabetic retinopathy ranging from 51·0% to 85·9% and specificity ranging from 60·4% to 83·7%.[3] Secondly, the results were inferior to those reported by the vendors from their in-house evaluations. This study could have created a more profound impact in diabetic eye screening had it openly identified the vendors. However, at the request of the vendors, the results were anonymised. Thus, we have been left in the unenviable position of knowing that some of these technologies are worse than others and yield data worse than the publicly presented data, but the exact technologies remain anonymised. This situation is metaphorically similar to being presented with the results of a trial for a cancer drug, with one drug showing a survival benefit, but not knowing the exact name of the drug for commercial reasons.

In this Comment, we argue that openly reported, independent, head-to-head validation studies should become the norm for evaluating the performance of AI-based health technologies, particularly for high value services such as screening. We describe the key issues encountered during such evaluations and illustrate them with the help of a concluded study in a national screening programme in the UK.

We recommend the following approach for independent head-to-head studies of AI-enabled medical devices. Firstly, the use case and setting should be defined (figure). Secondly, regulator-approved AI systems with compatible intended use statements[4] should be systematically identified to minimise selection bias. To facilitate this identification, medical device regulators should provide a searchable database of AI devices, enabling the easy identification of eligible devices; the US Food and Drug Administration is already on the forefront, but the search functionality is still limited. Thirdly, the dataset should be sufficiently large and representative—with labelling applied using a standardised and robust approach—and should not have been previously used in the development of any AI systems being evaluated. In most disease settings, such as proliferative diabetic retinopathy or aggressive cancers, some crucial features are particularly important to recognise; the dataset should include an adequate number of such critical cases to provide sufficiently precise estimates of accuracy but should also remain representative of the population of interest to enable, for example, cost-effectiveness analyses. Further, the studies should be strengthened by including sufficient numbers of critical cases from important population subgroups, such as demographic groups with known poor health outcomes. Fourthly, the performance that would meet or exceed the current standard for clinical deployment in the specific context should be predefined. This step will involve making decisions regarding the appropriate balance between sensitivity and specificity, which will vary across populations, countries, programmes, and health conditions. Some critical cases might warrant a high sensitivity, which should also be predefined. The registration of these studies, in the same manner as that for clinical trials, with all predefined outcomes stated, should become standard practice. Fifth, only companies that agree to openly publish their results

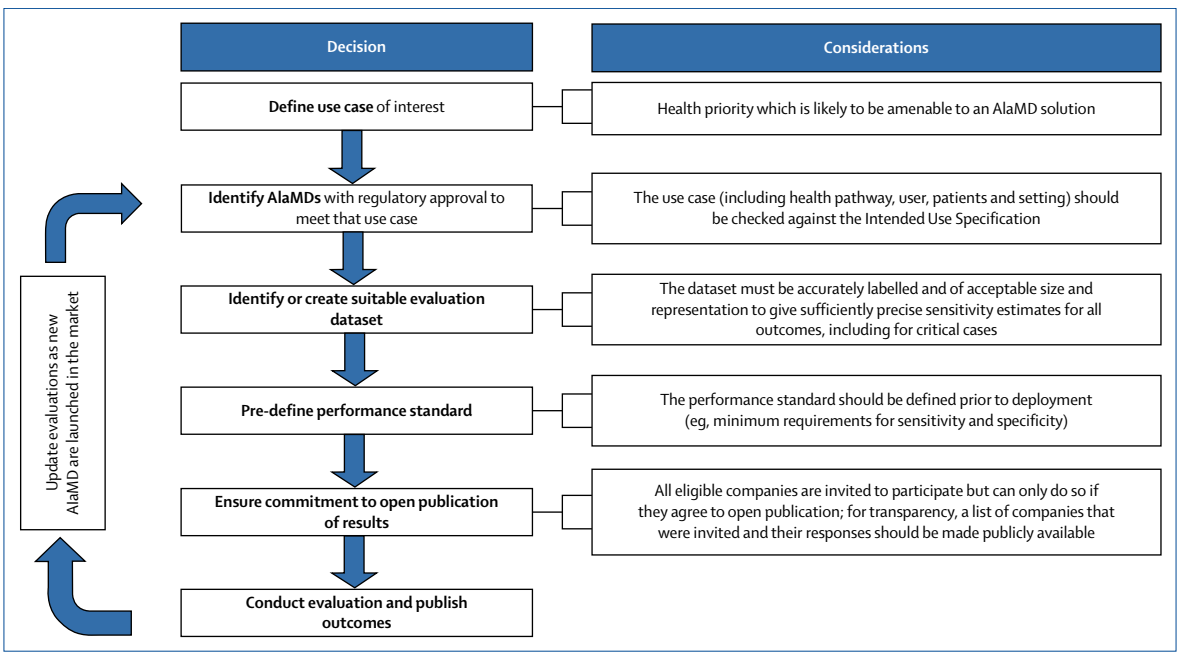

Figure: **Flow chart summarising the recommended approach for independent head-to-head studies of AI systems**
AIaMD=artificial intelligence as a medical device.

in a peer-reviewed journal, with the vendors named in the publication, should be provided with the opportunity to participate, acknowledging the fact that doing so might introduce self-selection bias with respect to vendors who agree to participate and those who do not agree. For transparency, the list of all companies that were invited to participate should be published. Vendor information could be anonymised during the peer review process to avoid any potential conflicts of interest during peer review. Additionally, as new vendors or versions of systems are introduced, the stakeholders (researchers, policy makers, and vendors) should be given the opportunity to repeat head-to-head evaluations, independent of any commercial interests, thus mitigating any disadvantage for new devices and encouraging innovation. Finally, the study results should be openly published with all the vendors being named.

To support UK health-care commissioners' decisions on which AI device(s) meet the standard for clinical deployment in the English Diabetic Eye Screening Programme and implementing the lessons learnt from the previous study by Lee and colleagues,[3] Alicja R Rudnicka and colleagues,[5] on behalf of the UK Artificial Intelligence and Automated Retinal Image Analysis Systems (ARIAS) Research Group, conducted a vendor-independent head-to-head study, published in *The Lancet Digital Health*, which compared the performance of different AI devices for diabetic retinopathy screening; the results have been openly published[5] and shared with commissioners. This head-to-head study used an ethnically diverse dataset of retinal images collected in the English NHS Diabetic Eye Screening Programme graded for referable diabetic retinopathy using a robust and standardised methodology.[6]

Although we acknowledge the commercial risk to companies who already have data that supports their claim to undergo a separate evaluation, these independent studies would provide greater credibility for any claims, enabling reliable and robust comparisons of performance among companies, potentially providing valuable insights into test accuracy that would otherwise not be accessible to vendors, and highlighting areas in which algorithms could benefit from improvement. The absence of such independent studies could put the industry, patients, and the overall trust in AI for health care at risk. Companies marketing AI devices should recognise this risk and be willing to contribute to the safe and responsible deployment of AI-assisted medical devices through engagement with transparent and independent evaluations. This open, head-to-head comparative evaluation[5] of different AI systems provides UK commissioners with key data required to make decisions regarding procurement and deployment within the English Diabetic Eye Screening Programme.

In summary, we propose that independent head-to-head comparative validation studies of AI-assisted medical devices are a necessary step towards safe and responsible clinical AI deployment. Journals can support this step by mandating the open naming of AI devices included in research. Policy makers and health-care providers can encourage the open reporting of results by recommending independent head-to-head studies with vendors named when making procurement decisions. The procurement of an AI health technology by a national screening programme is a considerable investment for health systems and a major financial win for any company. These screening programmes should leverage the opportunity to ensure that the best possible independent evidence is available for decision making.

CE has received fees from Heidelberg Engineering, Inozyme Pharma, and Boehringer Ingelheim. AT has received fees from Annexon, Apellis, Bayer, Genentech, Iveric Bio, Novartis, Oxurion, and Roche. All other authors declare no competing interests. The views expressed in this article are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care. MJB is supported by the Wellcome Trust (207472/Z/17/Z). CRC, AT, CE, AKD, CB, and MJB were responsible for the conception of this work. CRC drafted the initial manuscript. All authors reviewed and approved the final manuscript.

*Charles R Cleland, Adnan Tufail, Catherine Egan, Xiaoxuan Liu, Alastair K Denniston, Alicja Rudnicka, Christopher G Owen, Covadonga Bascaran, Matthew J Burton
charles.cleland@lshtm.ac.uk

International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK (CRC, CB, MJB); Eye Department, Kilimanjaro Christian Medical Centre, Moshi, Tanzania (CRC); National Institute for Health and Care Research (NIHR) Biomedical Research Centre (BRC) for Ophthalmology, Moorfields Hospital London NHS Foundation Trust, and Institute of Ophthalmology, University College London, London, UK (AT, CE, MJB); University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK (XL, AKD); National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK (XL, AKD); Population Health Research Institute, City St George's, University of London, London, UK (AR, CGO)

1 Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021; **181:** 1065–70.
2 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; **15:** e1002683.
3 Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* 2021; **44:** 1168–75.
4 Determination of intended use for 510(k) devices - guidance for CDRH staff (update to K98-1). US Food and Drug Administration Center for Devices and Radiological Health. Dec 3, 2002. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/determination-intended-use-510k-devices-guidance-cdrh-staff-update-k98-1 (accessed Aug 19, 2024).
5 Rudnicka AR, Shakespeare R, Chambers R, et al. Automated retinal image analysis systems to triage for grading of diabetic retinopathy: a large-scale, open-label, national screening programme in England. *Lancet Digit Health* 2025; **7:** 100914.
6 Fajtl J, Welikala RA, Barman S, et al. Trustworthy evaluation of clinical AI for analysis of medical images in diverse populations. *NEJM AI* 2024; **1:** AIoa2400353.