DECIDE-AI: new reporting guidelines to bridge the development to implementation gap in clinical artificial intelligence

The DECIDE-AI Steering Group*

* *A list of authors and their affiliations appears at the end of the paper.*

Published: 01 February 2021

The DECIDE-AI Steering Group., Vasey, B., Clifton, D.A. *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* (2021).

DOI : 10.1038/s41591-021-01229-5

https://www.nature.com/articles/s41591-021-01229-5

As an increasing number of clinical decision support systems driven by artificial intelligence progress from development to implementation, better guidance on the reporting of human factors and early-stage clinical evaluation is needed.

Recent years have seen an exponential growth in the number of artificial intelligence (AI) algorithms published in the medical literature, yet clinical impact in terms of patient outcomes remains to be demonstrated. One likely explanation for this so-called AI chasm[1] is an overemphasis on the technical aspects of the proposed algorithms, and insufficient attention to the factors affecting the interaction with their human users. As clinicians occupy, and are likely to keep occupying, the central role in patient care, it is essential to focus the development and evaluation of AI-based clinical algorithms on their potential to augment rather than replace human intelligence. However, AI-based decision support systems pose unique challenges to the traditional medical decision-making process, such as their frequent lack of explainability (the so-called "black box" problem) or their tendency to sometimes produce unexpected results. Hence, bridging algorithm development to bedside application while keeping humans at the centre of the design and evaluation process is a complicated task, and current guidance is incomplete.

We make the case for a robust early and small-scale clinical evaluation stage, between the *in silico* algorithm development/validation (covered by the upcoming TRIPOD-AI[2] and STARD-AI[3] statements) and large-scale clinical trials evaluating AI interventions (covered by the CONSORT-AI[4] statement). This step can be compared to a phase I/II trial for drug development or (a much closer analogy given the relationship between users' characteristics and the intervention's effectiveness) IDEAL stage IIa/IIb for surgical innovation.[5–7] Four key arguments support the need for this intermediary development stage, and its adequate reporting.

Human decision-making processes are complex and subject to many biases. It cannot be expected, even in the case of directive models, that human users will exactly follow all of the algorithm recommendations, especially if these users remain accountable for their decisions.[8] In order to accurately evaluate an algorithm's performance and avoid the research waste of conducting expensive large-scale trials with decision support systems whose interaction with human users is inadequate, it is essential to assess the actual impact of an algorithm on its users' decisions at an early stage. Additionally, consideration should be given to the difference between the development and target patient population, ensuring the algorithm's relevance in the implementation settings. Therefore, the assisted human performance and algorithm usability (not merely the algorithm's stand-alone outputs) need to be evaluated in the target clinical environment and reported as outcomes.

Because it cannot be assumed that users' decisions will mirror the algorithm's recommendations, it is also crucially important to test the safety profile of new algorithms not only in silico, but when used to influence human decisions. Skipping this step and moving directly forward to large scale-trials would expose a considerable number of patients to an unknown risk of harm, which is ethically unacceptable. Suboptimal safety standards have led to disastrous consequences in the early days of pharmacological trials; there is no need to repeat these mistakes with clinical AI.

Human factors (ergonomics) evaluation should happen as early as possible and needs iterative evaluation-design cycles. Technical requirements often evolve as a system starts being used, and users' expectations of a system also vary in the initial exposure period. For example, users might wish for an additional key variable to make sense of the algorithm recommendations, which in turn will require developers to access a totally different section of the electronic patient record. From an economic viewpoint, the sooner human factors evaluation occurs, the more cost effective it is likely to be. Finally, iterative design modification is difficult and inappropriate during large-scale trials causing a serious risk of invalidating the summative evaluation's conclusions, as the intervention tested is likely to have changed during trial. Early formative evaluation and rapid prototyping are therefore essential prior to large-scale trials.

Large-scale clinical trials are complex and expensive endeavours requiring careful preparation. A well-thought-out design is essential to produce valid and meaningful conclusions and needs background information about the intervention under evaluation. Not all such background information can be inferred from *in silico* evaluation and some data have to be collected in small-scale prospective studies. For example, the most appropriate outcomes for the trial, the expected effect size, the optimal inclusion and exclusion criteria for the user population, the evolution of the users' trust in the algorithm, and the most appropriate timing of decision support are crucial information which should be known to the investigators at the time of drafting trial protocols, and these could be derived from early formative evaluation. Other important considerations, such as how to best use the output of the algorithm or how this output is to be communicated to the patients, could also be investigated at this stage.

We believe that clear and transparent reporting on these aspects will not only avoid preventable harm and research waste, but also play a key role in transforming AI from a promising technology to an evidence-based component of modern medicine. This is why we have started a Delphi process[9,10] to reach expert consensus on the key information items that should be reported during the Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence (DECIDE-AI). The creation of the DECIDE-AI guidelines will be an open and transparent process and we will welcome expressions of interest from experts wishing to contribute.

References

1. A. Keane, P. & J. Topol, E. *npj Digital Medicine* 1, (2018).
2. Collins, G. S. & Moons, K. G. M. *Lancet* 393, 1577–1579 (2019).
3. Sounderajah, V. *et al. Nat. Med.* 26, 807–808 (2020).
4. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. *BMJ* 370, m3164 (2020).
5. McCulloch, P. *et al. Lancet* 374, 1105–1112 (2009).
6. Hirst, A. *et al Ann. Surg.* 269, 211–220 (2019).
7. Bilbro, N. A. *et al. Ann. Surg.* Publish Ah, (9000).
8. Price, W. N. 2nd, Gerke, S. & Cohen, I. G. *JAMA* (2019). doi:10.1001/jama.2019.15064
9. Dalkey, N. & Helmer, O. *Manage. Sci.* 9, 458–467 (1963).
10. Powell, C. *J. Adv. Nurs.* 41, 376–382 (2003).

Competing interests

XL is Industry Fellow (observer) with Hardian Health (No financial interest); DC declares academic grants from GlaxoSmithKline and personal fees from Oxford University Innovation, Biobeats, and Sensyne Health, outside the context of this work; BG is CEO and Founder of Healthplus.ai and declares consulting fees from NLC Ventures Netherlands; LM is director and owner of Morgan Human Systems Ltd; PW reports grants from National Institute for Health Research, grants from Wellcome, grants from Sensyne Health, personal fees from Sensyne Health, outside the submitted work.

Consortium

The DECIDE-AI Steering Group

Baptiste Vasey[1§], David A. Clifton[2], Gary S. Collins[3,4], Alastair K. Denniston[5], Livia Faes[6,7], Bart F. Geerts[8], Xiaoxuan Liu[5,7], Lauren Morgan[9], Peter Watkinson[10], Peter McCulloch[1]

[1] *Nuffield Department of Surgical Sciences, University of Oxford, UK*
[2] *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK*
[3] *Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[4] *NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK*
[5] *University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK*
[6] *Eye Clinic, Cantonal Hospital Lucerne, Lucerne, Switzerland*
[7] *Moorfields Eye Hospital NHS Foundation Trust, London, UK*
[8] *Healthplus.ai B.V., Amsterdam, The Netherlands*
[9] *Morgan Human Systems Ltd, Shrewsbury, UK*
[10] *Critical Care Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, UK*

[§] corresponding author: baptiste.vasey@nds.ox.ac.uk