

# Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence (DECIDE-AI)



## Project's scope and objectives

The present project aims to develop new reporting guidelines to guide early-stage clinical evaluation of clinical decision support systems (CDSS) based on artificial intelligence (AI) with the aim of making it more consistent, comprehensive and reproducible. A robust early formative evaluation, with an emphasis on the human users, is a necessary bridge from algorithm development to implementation and paves the way toward large-scale summative evaluation. The objective of DECIDE-AI is to improve reporting along four main axes:

- the performance of the algorithm when first used with humans in small-scale, actual clinical settings,
- the safety profile of the algorithm prior to large-scale utilisation,
- the human factors (ergonomic) evaluation of the algorithm,
- the preparatory steps towards large-scale (randomised controlled) clinical trials.

The aim of the DECIDE-AI guidelines<sup>1</sup> is to bridge the gap between the TRIPOD-AI<sup>2</sup> and STARD-AI<sup>3</sup> (algorithm development/validation), and CONSORT/SPIRIT-AI<sup>4,5</sup> (large-scale summative evaluation) statements. This necessary intermediary step can be compared to pharma phase I/II clinical trials or (a closer analogy given the relationship between users' characteristics and the intervention's effectiveness) IDEAL stage IIa/IIb studies for surgical innovation (figure 1).<sup>6-8</sup>

## Background

Over the recent years, the number of publications describing AI-based clinical algorithms has grown exponentially in the medical literature, yet actual clinical impact in terms of patient outcomes remains to be demonstrated. One likely reason for this so-called AI chasm<sup>9</sup> is an overemphasis on the technical aspects of the algorithms, leading to insufficient attention to the factors surrounding the human-computer interaction. Because clinicians occupy, and will likely maintain, a 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. Progressing from the *in silico* development (i.e. through computer simulation) of an algorithm to its clinical application while keeping humans at the centre of the design and evaluation process is not trivial, though, and current guidance is incomplete.

A systematic review on the impact of AI-based diagnostic CDSS on clinician performance found that the approach to early-stage clinical evaluation was very heterogeneous, the risk of bias in most included studies was high, and almost no consideration was given to human factors.<sup>10</sup> The review also demonstrated that users almost always decided to override some of the algorithm's recommendations, thereby highlighting the need to evaluate the assisted human performance (not merely the algorithm's stand-alone outputs) in the, clinical environment and to report it as the primary outcome.

Inconsistent and low-quality evaluation of AI-based algorithms not only puts patients at risk of errors but is also detrimental to the acceptance of this technology by healthcare professionals, and therefore to its deployment in clinical settings. Regulatory bodies are currently refining their strategies on how to best to assess this emerging technology and more academic inputs are needed to inform the discussion around this process.<sup>11</sup>

## Project justification

When comparing the development pathway of AI-based algorithms with those of new drugs or surgical procedures, the need for a comprehensive and robust early-stage clinical evaluation becomes even clearer. Like a new molecule, the safety profile of a new algorithm has to be assessed with a small number of real users before being deployed to a larger group. Like a surgical intervention, the effectiveness of a CDSS is inherently linked to its operators and evolves with time as the operators become accustomed to using it. To ensure safety and avoid research waste, four important aspects of a new algorithm need to be evaluated at small-scale before progressing to larger, ideally multicentric, summative evaluation studies.

**1. the algorithm performance when first used with humans**, in actual clinical settings. Human users don't always follow the algorithm recommendations. An algorithm which proved very promising in theory can perform poorly when used with humans in actual clinical settings (actual disease prevalence, access to information unknown to the algorithm, etc.). Large-scale clinical trials are expensive and time-consuming; it is important to ensure that only the best performing algorithms progress to this advanced stage of evaluation.

**2. the algorithm's safety profile.** As with drugs or surgical interventions, it would be reckless (if not unethical) to directly roll out a new technology, untested with humans, in an extended population of patients in the context of a large-scale trial. The safety of an algorithm needs to be first evaluated on a small cohort of patients with the appropriate surveillance.

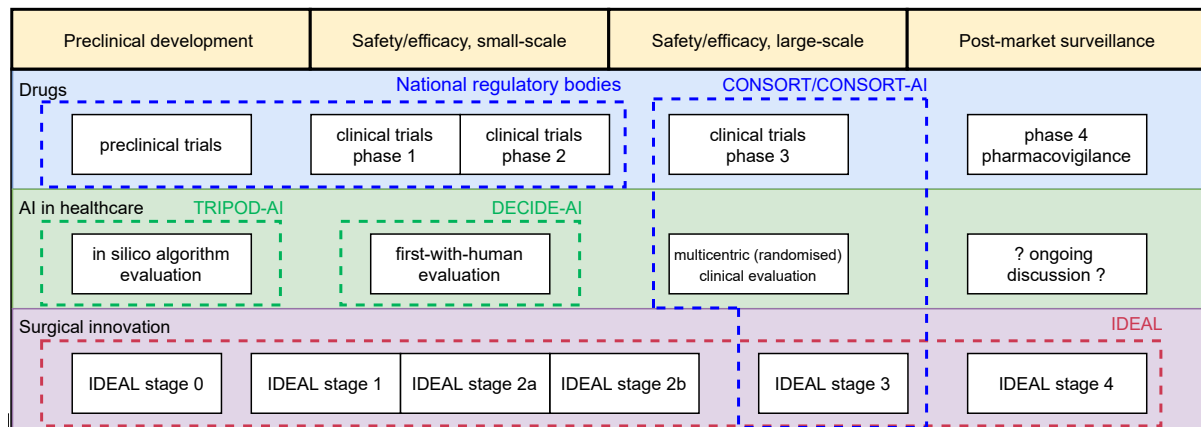
**3. the human factors (ergonomic) influencing the use of the algorithm.** The design of a tool or process and the way it is integrated into a workflow have a strong influence on its users' physical and cognitive performance. Optimising the ergonomics of a CDSS in early stages can benefit later performance and is essential for acceptance in clinical settings during trials. Early-stage human factors evaluation may require several design-measurement iterative cycles, which are impossible to achieve during large-scale trials. There is also a strong economic argument that the sooner human factors are considered, the more cost effective it is likely to be.

**4. the preparation for a subsequent large-scale summative evaluation.** Large clinical trials are complex and expensive endeavours, which need careful planning and preliminary data to define their main parameters. Smaller-scale evaluations are indispensable opportunities to explore key variables, like learning curves or effect size for example, and gain important information for the design of trial protocols. Such preparation increases the chances of success and helps to reduce research waste.

At present, however, there is no accepted guidance for the reporting of this crucial intermediary phase of development. This is the gap DECIDE-AI aims to address.

## Methodology

A Delphi process<sup>12,13</sup> comprising two rounds and a consensus meeting will be used to reach expert consensus. The first round will collect, through open-ended questions, experts' opinions on the necessary reporting items. It will also contain a scoring exercise, where participants will have the opportunity to provide feedback on a list of preliminary items developed by the DECIDE-AI steering group and based on systematic reviews, academic publications, regulatory documentation and institutional frameworks related to AI implementation and evaluation.<sup>14-23</sup> In the second round, participants will score each item of an updated item list. A consensus meeting with a selected subset of experts will discuss the results of the two rounds and select the final item list, based on the outcome of the second scoring exercise.



The dotted lines indicate reporting guidelines.

## Steering Group

- Prof David Clifton, Computational Health Informatics Lab, University of Oxford
- Prof Gary Collins, TRIPOD and UK EQUATOR network, University of Oxford
- Prof Alastair Denniston, CONSORT/SPIRIT-AI, University of Birmingham
- Dr Livia Faes, CONSORT/SPIRIT-AI, Moorfields Eye Hospital
- Dr Bart Geerts, CEO and founder of healthplus.ai, University of Amsterdam
- Dr Xiaoxuan Liu, CONSORT/SPIRIT-AI, University of Birmingham
- Dr Piyush Mathur, Department of General Anesthesiology, Cleveland Clinic, Ohio
- Prof Peter McCulloch, Chair of the IDEAL collaboration, University of Oxford
- Dr Lauren Morgan, human factors specialist, Morgan Human Systems Ltd
- Dr Suchi Saria, Department of Computer Science, Johns Hopkins University
- Baptiste Vasey, Nuffield Department of Surgical Sciences, University of Oxford
- Dr Peter Watkinson, Critical Care Research Group, University of Oxford

## Reference list

1. Vasey, B. *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
2. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* 393, 1577–1579 (2019).
3. Sounderajah, V. *et al.* Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat. Med.* 26, 807–808 (2020).
4. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 370, m3164 (2020).
5. Cruz Rivera, S. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* 26, 1351–1363 (2020).
6. McCulloch, P. *et al.* No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* 374, 1105–1112 (2009).
7. Hirst, A. *et al.* No Surgical Innovation Without Evaluation: Evolution and Further Development of the IDEAL Framework and Recommendations. *Ann. Surg.* 269, 211–220 (2019).
8. Bilbro, N. A. *et al.* The IDEAL Reporting Guidelines: A Delphi Consensus Statement Stage specific recommendations for reporting the evaluation of surgical innovation. *Ann. Surg.* Publish Ah, (9000).
9. A. Keane, P. & J. Topol, E. *With an eye to AI and autonomous diagnosis.* *npj Digital Medicine* 1, (2018).
10. Vasey, B. *et al.* Association of Clinician Diagnostic Performance With Machine Learning–Based Decision Support Systems: A Systematic Review. *JAMA Netw. Open* 4, e211276–e211276 (2021).
11. Smith, J. A. *et al.* Financial interests and evidence in public comments on the FDA framework for modifications to artificial intelligence/machine learning–based medical devices. *medRxiv* 2019.12.11.19013953 (2019). doi:10.1101/2019.12.11.19013953
12. Dalkey, N. & Helmer, O. An Experimental Application of the DELPHI Method to the Use of Experts. *Manage. Sci.* 9, 458–467 (1963).
13. Powell, C. The Delphi technique: myths and realities. *J. Adv. Nurs.* 41, 376–382 (2003).
14. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368, m689 (2020).
15. Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci. Eng. Ethics* (2019). doi:10.1007/s11948-019-00165-5
16. Vollmer, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 368, l6927 (2020).
17. IMDRF Software as Medical Device (SaMD) Working Group. ‘Software as a Medical Device’: Possible Framework for Risk Categorization and Corresponding Considerations. (2014).
18. IMDRF Software as Medical Device (SaMD) Working Group. *Software as a Medical Device (SaMD): Clinical Evaluation.* (2017).
19. National Institute for Health and Care Excellence (NICE). *Evidence standards framework for digital health technologies.* (2019).
20. Accelerated Access Collaborative & NHSx. *AI-Award Evaluation Playbook – Version 1.* (2020).
21. Independent High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI.* (2019).
22. Xie, Y. *et al.* Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Transl. Vis. Sci. Technol.* 9, 22 (2020).
23. Sujan, M. *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Heal. & Care Informatics* 26, e100081 (2019).