

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 the early-stage clinical evaluation of clinical decision support systems (CDSS) based on artificial intelligence (AI).¹

Reporting guidelines are a set of items on which authors should report when writing up a manuscript. One such item could be “report on the number of patients enrolled and their baseline characteristics”. To follow the guidelines, authors should then clearly state the number of patients enrolled and describe the patient population (age, sex, comorbidities, etc). The goal of such guidelines is to make sure authors report on the key information necessary for readers to make an informed appraisal of the studies’ conclusions. To follow up on the same example, the interpretation of a study will be very different if the results are based on a group of 10 or of 1000 patients, and readers need to have this information. In the field of clinical AI, other guidelines, like CONSORT-AI,² have already been developed (or are in development) to address other stages of technology evaluation.

Clinical decision support systems are applications used to inform, support or influence clinician decision making, most of the time for direct patient care. This can for example be an application which calculates the risk of complication following a specific surgery and will then give information pro or con the said surgery. An important aspect of such systems is that their performance is strongly dependent from the humans using them. Even the best possible CDSS has no value if the clinicians using it don’t trust its suggestions.

Artificial intelligence is a loosely defined concept referring to an algorithm capable of performing a task which would normally require human intelligence to complete. For example, picking a clinician-selected drug in a shelf does not require artificial intelligence (the software engineer would have coded the location of each drug in the shelf and the computer would just match a label and retrieve the drug). However, selecting the appropriate drug for a given patient requires the integration of various sources of information, the weighting of risk or patient preferences and would require intelligence. True artificial intelligence does not exist yet. However, machine learning (computers able to learn from data information unknown to their coder) has made impressive progress in the past year and now compete with human clinicians for the specific task they have been trained in (for example, detecting breast cancer on a mammogram). For harmonisation of terminology, we will only refer to artificial intelligence in this document.

The development pathway of medical technology varies depending on what is evaluated. The development of drugs is now largely formalised and similar guidance also exists for surgical innovation.^{3,4} DECIDE-AI draws experience from these other fields of medicine and focuses on the early clinical evaluation of AI-based CDSS. This is the stage in which CDSS, which have been successfully validated through computer simulation, are used for the first time in actual clinical settings. Of course, this early, small-scale evaluation will not be enough to prove the impact of the CDSS on patient care (this is done later through large-scale clinical trials), but this stage of evaluation is especially important for several reasons, that DECIDE-AI aims to address (see figure 1):

1. **Confirm the algorithm performance when 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 (time pressure from the environment, clinician having access to information unknown to the CDSS, etc.).
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 (ergonomics) influencing the use of the algorithm**. Human factors are the environmental, organisational, mental and physical factors that influence an individual's performance. 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-evaluation-modification cycles. In other words, the CDSS is given to clinicians, the clinicians will notice problems in the use of the CDSS, report them and the engineers will modify the CDSS design to resolve these problems. The new prototype will then be given to clinicians again and a new cycle starts.
4. **the preparation for a subsequent large-scale evaluation**. Large clinical trials are necessary to prove the effectiveness of a CDSS in practice, but they are complex and expensive. They need careful planning and preliminary data to define their main parameters (for example how many patients should be included, how long should the clinicians be trained with the CDSS before starting using it, etc). Smaller-scale evaluations are indispensable opportunities to explore these key variables 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^{5,6} comprising two rounds and a consensus meeting will be used to reach expert consensus. This is a methodology developed to progressively bring expert to agree on a set of final recommendations. 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 reporting 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.⁷⁻¹⁶ Between the two rounds, the research team will update the provisory items list to include the participants' inputs. In the second round, participants will score each item of an updated item list. Through this process, more and less "popular" items will emerge. A consensus meeting with a selected subset of experts will finally discuss the results of the two rounds. The consensus group will select the final items list, based on the outcome of the scoring exercises and the group discussion. We aim to invite several patient representatives to the Delphi rounds and two to the consensus meeting.

How can you contribute?

We would like to hear your opinion on what should, from a patient perspective, be reported when authors present the results of the early-stage evaluation of an AI-based CDSS. In other words, what information would patient like to know about a CDSS before accepting their care being influenced by it? The questions and items in the Delphi's first round can be quite technical sometimes, but there are no right or wrong answers. What really interests us is your opinion. Moreover, you will always have the option to answer "I don't know", an option that many of the other experts have used already.

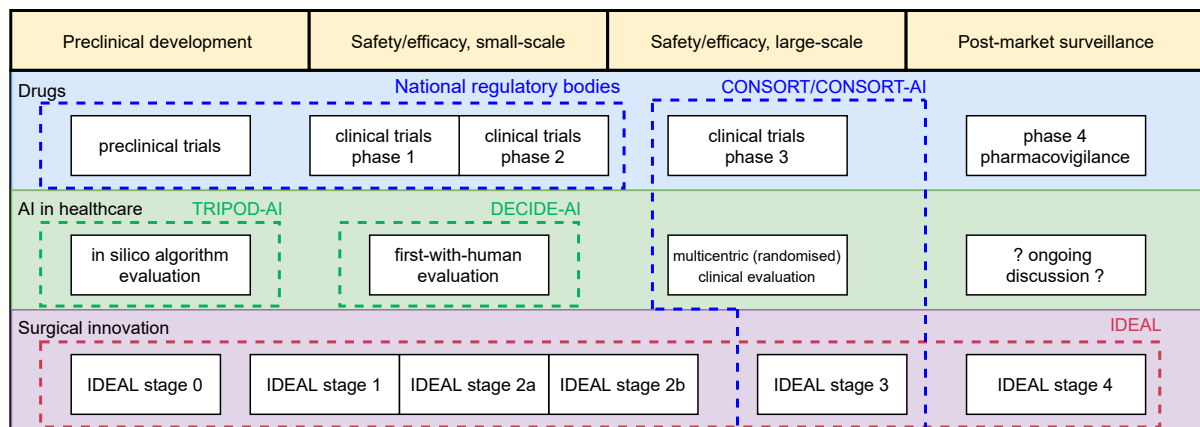


Figure 1: Comparison of the development pathways for drugs, AI-based algorithms and surgical innovation. The dotted lines indicate reporting guidelines.

Steering Group

- Prof David Clifton, Professor of clinical machine learning, 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, Cleveland Clinic, Ohio
- Prof Peter McCulloch, Chair of the IDEAL collaboration, University of Oxford
- Dr Lauren Morgan, human factors specialist, Morgan Human Systems Ltd
- 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. 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).
3. McCulloch, P. *et al.* No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* 374, 1105–1112 (2009).
4. 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).
5. Dalkey, N. & Helmer, O. An Experimental Application of the DELPHI Method to the Use of Experts. *Manage. Sci.* 9, 458–467 (1963).
6. Powell, C. The Delphi technique: myths and realities. *J. Adv. Nurs.* 41, 376–382 (2003).
7. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368, m689 (2020).
8. 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
9. 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).
10. IMDRF Software as Medical Device (SaMD) Working Group. *'Software as a Medical Device': Possible Framework for Risk Categorization and Corresponding Considerations.* (2014).
11. IMDRF Software as Medical Device (SaMD) Working Group. *Software as a Medical Device (SaMD): Clinical Evaluation.* (2017).
12. National Institute for Health and Care Excellence (NICE). *Evidence standards framework for digital health technologies.* (2019).
13. Accelerated Access Collaborative & NHSx. *AI-Award Evaluation Playbook – Version 1.* (2020).
14. Independent High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI.* (2019).
15. Xie, Y. *et al.* Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Transl. Vis. Sci. Technol.* 9, 22 (2020).
16. Sujan, M. *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Heal. & Care Informatics* 26, e100081 (2019).