BMJ Surgery, Interventions, & Health Technologies

Addressing the reporting chasm of artificial intelligence research: the DECIDE-AI reporting guidelines

John Gerrard Hanrahan ,[1,2] Danyal Zaman Khan ,[1,2] Hani J Marcus [1,2]

To cite: Hanrahan JG, Khan DZ, Marcus HJ. Addressing the reporting chasm of artificial intelligence research: the DECIDE-AI reporting guidelines. *BMJ Surg Interv Health Technologies* 2022;**4**:e000154. doi:10.1136/bmjsit-2022-000154

Received 04 May 2022
Accepted 01 June 2022

[1]Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK
[2]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, UK

Correspondence to
Dr John Gerrard Hanrahan; j.hanrahan@ucl.ac.uk

EDITORIAL

The meteoric rise of artificial intelligence (AI) to the forefront of healthcare innovation has unearthed an array of avenues for surgical researchers to pursue. Applications found throughout the surgical patient pathway mean AI offers new-found support systems for clinical decision-making. Indeed, a growing number of technologies are entering clinical practice,[1] with a recent review evaluating randomised controlled trials of diagnostic prediction tools suggests that potential benefits of AI that contemporary healthcare stands to realise.

However, the pathway to translation to the bedside for these technologies is variable. Captured aptly in a recent editorial, there are clear examples of AI technologies already approved for clinical use in the USA, both with and without evaluation through randomised controlled trials.[2] This speaks to a wider problem of evaluation in AI innovation, where insufficient reporting in randomised controlled trials prompted the development of several reporting guidelines, examples including the Consolidated Standards of Reporting Trials-AI and Standard Protocol Items: Recommendations for Interventional Trials-AI guidelines advising the minimum reporting standards for clinical trials and protocols, respectively. Similarly, guidance for the initial stages of AI development has been developed, namely, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD-AI) guidelines for machine learning (ML) prediction models.[3]

Yet, when one looks at the process of AI translation, from in silico to clinical trial, an evaluation chasm becomes obvious, with guidance lacking on studies reflecting stages 2a and 2b of the IDEAL (Idea, Development, Exploration, Assessment, Long-term study) collaborative. These stages reflect the refinement and preparation for larger clinical studies, which are influenced by factors from the operator including learning curves or training; the health system the technologies enter into or organisational factors such as integration into clinical workflows. Study design features such as patient selection for both training and testing an intervention, and even the AI model itself, are crucial factors to consider prior to large-scale testing.

Vasey and colleagues have identified a gap in the reporting guidelines for evaluating AI-driven decision support systems, producing reporting guidelines to support the evaluation of their early stages. This was achieved through an international, two-round modified Delphi consensus process producing a 17 AI-specific item and 10 generic item reporting guidelines (DECIDE-AI), informing the reporting of early-stage clinical studies of AI-based decision support systems in healthcare.

The systems perspective taken by Vasey *et al* frame AI decision-support systems as complex interventions.[4] This perspective clearly elucidates the importance of understanding of the workflow or clinical process interventions are intended to enter, alongside the evaluation setting of the AI. Reporting of such demonstrates the setting, or even system-specific evaluation in the selected trial which may be important in judging intervention efficacy when applied to the same clinical problem in alternate health systems or settings.

Furthermore, the emulation of aviation or military human factors appraisal is another value of the DECIDE-AI guidelines, particularly as the augmentative nature of AI decision-support systems rely on human-computer interactions. It is evident, for example, in surgery that learning-curves of surgeons influence clinical outcomes,[5] meaning complex interventions including AI-based tools must account for this during

the evaluation process. Failing to do so may result in intervention failure in larger clinical trials, at cost to researchers and developers, but with perhaps greater cost to trial participants.

Considering these factors, rigorously and systematically, are undoubted means of improving translation of AI interventions from bench-to-bedside. The vitality of which is two-fold; pursuit of evidence-based medicine principles for safe evaluation of a technology, testing and developing them in real-world health systems, coupled with more accurate determination of efficacy and effectiveness, progressing evaluation towards more realistic settings.

One cannot claim perfection when deciding on reporting guidelines, and Vasey and colleagues recognise the known limitations as they achieved consensus from their spectrum of experts. Yet, it is clear that they have provided a robust foundation to foster systematic and transparent reporting to guide the early-stage clinical evaluation of AI technologies. Recognition and improvement of the translation processes' weaknesses certainly stand to aid AI innovators of tomorrow, with clinical dividends to follow.

Twitter John Gerrard Hanrahan @johnhanrahan1 and Hani J Marcus @Hani_marcus

Contributors JGH drafted the editorial. DZZK assisted the draft, provided substantial edits and manuscript review. HJM is the senior author invited to provide the editorial, who substantially edited and reviewed the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iDs
John Gerrard Hanrahan http://orcid.org/0000-0003-4584-2298
Danyal Zaman Khan http://orcid.org/0000-0001-9213-2550
Hani J Marcus http://orcid.org/0000-0001-8000-392X

REFERENCES

1 Zhou Q, Chen Z, Cao Y. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. npj digit. *Med* 2021;4:154.
2 Angus DC. Randomized clinical trials of artificial intelligence. *JAMA* 2020;323:1043–5.
3 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
4 DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186–7.
5 van Workum F, Stenstra MHBC, Berkelmans GHK, *et al*. Learning curve and associated morbidity of minimally invasive esophagectomy: a retrospective multicenter study. *Ann Surg* 2019;269:88–94.