[1] Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

[2] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[3] EPI-Centre, KU Leuven, Leuven, Belgium

[4] Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl https://orcid.org/0000-0002-3037-122X

Cite this as: *BMJ* 2020;370:m3505

http://dx.doi.org/10.1136/bmj.m3505

Published: 09 September 2020

Demystifying AI in healthcare

Well conducted and transparently reported trials would be an excellent start

Laure Wynants, [1,2,3] Luc J M Smits, [1] Ben Van Calster[2,3,4]

In academia and society at large, attention on artificial intelligence (AI) in healthcare is tremendous. Although many researchers and commentators claim that AI improves screening, diagnosis, and prognostication, those who delve deeper will notice a scarcity of external validation studies and randomised controlled trials evaluating the true impact of AI on healthcare.[1-3] Findings from the few published randomised controlled trials are mixed. In one trial, endoscopy assisted by an automatic AI detection system found more colorectal adenomas than did unassisted endoscopy.[4] In another, an AI platform for diagnosing childhood cataracts was less accurate than a senior consultant.[5] To gauge the quality of such evidence, readers need a detailed account of study methods and results. Systematic reviews, however, show that studies on AI are often poorly reported.[2 6]

Reporting guidelines

New extensions of the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) (doi:10.1136/bmj.m3210) and CONSORT (Consolidated Standards of Reporting Trials) (doi:10.1136/bmj.m3164) reporting guidelines, published in *The BMJ*, encourage authors to be transparent and comprehensive when writing protocols for trials that evaluate AI interventions,[7] and when reporting the results of such trials.[8] They cover important issues specific to AI interventions, such as specifying the level of expertise required for researchers interacting with the study's AI (for example, to identify a region of interest on an image, or to translate AI output into clinical decisions). The operational requirements for integrating AI into the study's clinical setting also must be clear, as well as any need to fine tune an AI algorithm using data from the local environment.

We can anticipate a positive effect of these reporting guidelines on the quality (and perhaps quantity) of trial reports in this rapidly developing area. Registering a trial protocol improves transparency and discourages research practices that might yield misleading results, such as switching the primary outcome after the results are known.[9] Similarly, empirical research suggests that CONSORT guidelines improved the quality of reporting, but that it remains suboptimal.[10 11] Funders, scientific publishers, and peer reviewers have an important responsibility to enforce protocol registration and the adoption of appropriate guidelines.[11]

But even a transparently reported study can lead to misguided conclusions if the trial is poorly designed, if it targets an inappropriate primary outcome, or if the AI system is not well embedded in the clinician's digital environment and workflow. In addition, owing to the difficulty and cost of running randomised controlled trials, it is important to evaluate the performance of AI algorithms in external validation studies first.[1-3]

One example of a primary outcome that could lead to unjustified claims about AI's benefits is the number of detected cases in a trial comparing clinicians' diagnostic performance with or without AI support. Such a trial is likely to show that AI helps detect more cases, even if the AI's alerts are completely random. A balanced evaluation must weigh up the increase in detected cases against the risk of false alerts.

Another example of the potential for misleading results is a trial of a very accurate AI system that has poor user adherence as a result of the way it is embedded in the clinician's environment. Poor adherence might be an important reason why clinical decision support systems have largely failed to improve patient health or reduce healthcare costs in trials.[12 13] Factors that have been shown to improve outcomes associated with clinical decision support systems include user friendliness, involving stakeholders in implementation, and using systems that give actionable recommendations, nudge users to comply (for example, by asking for a reason to overrule a recommendation), and target clinicians and patients simultaneously in a shared decision making context.[12 13]

Reporting harm

Similar to the monitoring of drug side effects, AI errors and other associated harms must be monitored and reported—both during trials and later in clinical practice. The new CONSORT and SPIRIT extensions encourage transparent reporting of errors, such as errors in diagnosing rare tumour subtypes or diagnostic errors in certain population subgroups.

One particularly worrying type of error arises from underrepresentation of minorities in the training data for AI systems—such as an application for detecting melanoma that is trained only on white skin. Another is the replication of social biases such as delayed lung cancer diagnosis in patients of low socioeconomic status.[14 15] By mechanisms such as these, AI replicates and could even exacerbate health inequities. This is particularly harmful when an AI system is wrongly perceived as objective and free from bias. Using large and diverse samples that allow subgroup analyses provides an opportunity to tackle these problems.

Despite the above considerations, we have an exciting new era to look forward to, in which the true potential of AI will gradually emerge. Sceptics might become enthusiasts, enthusiasts might be disappointed. But whatever happens, well designed trials, registered and published protocols, and transparent reporting will help ensure that a nuanced appraisal of all AI interventions is based on robust evidence instead of fears or aspirations.

Competing interests: *The BMJ* has judged that there are no disqualifying financial ties to commercial companies. The authors declare the following other interests: LW and BVC acknowledge funding from Internal Funds KU Leuven, KOOR, and the COVID-19 Fund, and Research Foundation-Flanders (FWO), all unrelated to the current editorial. LW and LJMS acknowledge funding from ZonMw, unrelated to the current editorial.

Provenance and peer review: Commissioned; not peer reviewed.

1 Wiens J, Saria S, Sendak M, etal. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337-40. doi: 10.1038/s41591-019-0548-6 pmid: 31427808

2 Nagendran M, Chen Y, Lovejoy CA, etal. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689. doi: 10.1136/bmj.m689 pmid: 32213531

3 Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works?*J Am Med Inform Assoc* 2019;26:1651-4. doi: 10.1093/jamia/ocz130 pmid: 31373357

4 Wang P, Berzin TM, Glissen Brown JR, etal. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68:1813-9. doi: 10.1136/gutjnl-2018-317500 pmid: 30814121

5 Lin H, Li R, Liu Z, etal. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine* 2019;9:52-9. doi: 10.1016/j.eclinm.2019.03.001 pmid: 31143882

6 Wynants L, Van Calster B, Collins GS, etal. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;369:m1328. doi: 10.1136/bmj.m1328 pmid: 32265220

7 Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJSPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* 2020;370:m3210.

8 Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AKSPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020;370:m3164.

9 Odutayo A, Altman DG, Hopewell S, Shakir M, Hsiao AJ, Emdin CA. Reporting of a Publicly Accessible Protocol and Its Association With Positive Study Findings in Cardiovascular Trials (from the Epidemiological Study of Randomized Trials [ESORT]). *Am J Cardiol* 2015;116:1280-3. doi: 10.1016/j.amjcard.2015.07.046 pmid: 26282722

10 Moher D, Jones A, Lepage LCONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001;285:1992-5. doi: 10.1001/jama.285.15.1992 pmid: 11308436

11 Vassar M, Jellison S, Wendelbo H, Wayant C, Gray H, Bibens M.Using the CONSORT statement to evaluate the completeness of reporting of addiction randomised trials: a cross-sectional review. *BMJ Open* 2019;9:e032024.

12 Bright TJ, Wong A, Dhurjati R, etal. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med* 2012;157:29-43. doi: 10.7326/0003-4819-157-1-201207030-00450 pmid: 22751758

13 Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;330:765. doi: 10.1136/bmj.38398.500764.8F pmid: 15767266

14 Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA* 2019;322:2377-8. doi: 10.1001/jama.2019.18058 pmid: 31755905

15 Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med* 2018;169:866-72. doi: 10.7326/M18-1990 pmid: 30508424