Reporting guidelines for artificial intelligence in medical research

J. Peter Campbell, MD, MPH, Aaron Y Lee, MD, MSCI, Michael Abràmoff, MD, Pearse A. Keane, MD, FRCOphth, Daniel SW Ting, MD PhD, and Michael F. Chiang, MD

Financial Support: JPC and MFC are supported by R01EY19474, R01EY031331, K12EY027720, and P30EY10572 from the National Institutes of Health (Bethesda, MD); and by unrestricted departmental funding and a Career Development Award from the Research to Prevent Blindness (JPC). AYL is supported by NIH/NEI K23EY029246, NIH P30EY10572, and an unrestricted grant from Research to Prevent Blindness. The sponsors / funding organizations had no role in the design or conduct of this research.

Financial Disclosures:

Michael D Abramoff, IDx (I, F,E,P,S), Alimera (F). J. Peter Campbell, Genentech(F). Aaron Y Lee, US FDA (E), Genentech (C), Topcon (C), Verana Health (C), Santen (F), Novartis (F), Carl Zeiss Meditec (F). Pearse A. Keane, DeepMind Technologies (C), Roche (C), Novartis (C), Apellis (C), Bayer (F), Allergan (F), Topcon (F), Heidelberg Engineering (F). Daniel Ting, EyRIS (IP), Novartis (C), Ocutrx (I, C), Optomed (C).

Corresponding Author:

Michael F Chiang
Address?
Contact?

Every so often a technology with the potential to disrupt clinical practice emerges and the medical literature explodes with new studies. These seismic events present a challenge to the peer review process, since many reviewers and editorial board members may be unfamiliar with how to evaluate them. Complicating matters, early adopters and thought leaders may not use consistent terminology, report results similarly, or fully appreciate the potential for inaccurate conclusions based on interpretation errors. There are thus two key motivations for developing reporting standards for academic research involving novel technologies. First, nonstandard reporting may limit the validity, comparability, and utility of research; standardization improves the return on investment of all research efforts. Second, clinical decisions based on unfamiliar technology may cause harm, either in the form of patient harm or inequity, either because the results are not valid in general, or are not generalizable to that patient in particular. In the first case, based on misinterpretations of data, we might come to conclusions that are not valid. In the second, we might come to valid conclusions based on the study data or population, but then apply them to datasets or other populations that differ in some meaningful, but often unknowable, way and come to incorrect conclusions as a result. As clinicians, due to our obligation to *primum non nocere* and more generally to ensure the bioethical principles of nonmaleficence and justice, it is incumbent on us to understand emerging technologies as they relate to our clinical care for patients.

We are currently in the midst of an epidemic of papers involving artificial intelligence (AI) algorithms in clinical medicine. Between 2015 and 2020 there were 728 publications using the terms artificial intelligence or deep learning and ophthalmology with 10 times as many in 2020 compared with 2017. As the Food and Drug Administration (FDA) develops new pathways for regulatory approval,[1] there is a growing appreciation not only for the potential benefits of AI in clinical medicine, but also the ways that it can fail and/or cause harm when implemented.[2] In order to facilitate the development of clinical AI devices that are not only efficacious in a research study, but safe, effective, equitable and reliable in practice, there is a relatively urgent need to standardize the reporting of AI papers by ensuring minimum necessary details for critical review, interpretation, and application of AI.

Originally designed for randomized clinical trials (RCTs), the CONSORT (Consolidated Standards of Reporting Trials) and the accompanying SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) guidelines were developed to standardize reporting of clinical trials and clinical trial protocols, respectively.[3,4] They have been widely adopted by medical journals, streamlining and standardizing the review process, enhancing the ability to compare between trials, and overall improved the interpretability of clinical trial results. There have been several "extensions" to the original guidelines addressing study designs that are not RCTs (http://www.consort-statement.org/extensions), including study designs from pilot and feasibility studies to herbal medicine intervention studies.[1,5] Over the last year, there has been an international effort to develop AI extensions both to CONSORT and SPIRIT, which are being simultaneously published in Nature Medicine, the British Medical Journal, and Lancet Digital Health.[6,7] The papers by Liu et al, and Rivera et al meticulously describe the process for developing AI specific guidelines that are considered essential for reporting AI clinical trials to be added to the existing CONSORT (14 additional) and SPIRIT (15 additional) guidelines. Until now, there has been no requirement for pre-registration, such as on clinicaltrials.gov, even though this has been shown to increase replication and lower effect size of studies compared to post hoc in- and exclusion as well as statistical analysis.[8] These AI specific guidelines fall into 3 general categories/concepts that are important to understand.

What is the device and what is it intended to do?

There are several specific recommendations that fall into this general category. First, to ensure transparency, the guidelines recommend specifying that the intervention involves AI within the title and/or abstract. Second, the methods need to specify exactly what was studied (hardware, software, version(s), etc), including internal thresholds. Third, the indication for use (IFU) needs to be explicitly defined, including by whom (who is the user) and where within the clinical pathway. For example, although published AI algorithms can both 1) detect referable diabetic retinopathy, and 2) specify the level of retinopathy, these are separately evaluable IFUs as they may each be utilized by different healthcare professionals and in different clinical practice settings. Although not specifically mentioned in the CONSORT-AI and SPIRIT-AI extensions, it is also important to consider the hierarchy of the truth to which the AI output is compared, from a single reader, to multiple readers, to an independent reading center. Ultimately, the most robust reference standards will be clinical outcomes, or outcomes that have been validated as equivalent to clinical outcomes.[9] Fourth, the input needs to be strictly defined including for imaging studies any technical requirements such as image quality, field of view, resolution, and camera device and model. Finally, the output should be in line with the IFU and its integration into the clinical care pathway defined and explained. Fundamentally, this set of guidelines is meant to ensure that the entire end-to-end pathway for the technology is reliable and reproducible when applied to a similar population. That is, at least for clinical trials, the unit of evaluation ought not be the algorithm, but the entire clinical pathway.

Who and what was studied?

In the same way that the results of a phase III clinical trial may not generalize to a population of patients that is dissimilar from those studied, the performance of an AI device is highly sensitive to the underlying population.[10] Thus, several of the AI extension guidelines relate to strictly defining the inclusion and exclusion criteria for who (which patients) and what (the type of data) that was studied. In addition, the methods should specify whether there was any human interaction involved in selecting which inputs were studied, and which were excluded. Specifically, the process for assessing and handling low quality data needs to be defined. In practical real world AI validation, these issues are critical, since many research datasets used for training are culled of low quality images and/or images that may not perfectly demonstrate a class label. If these difficult to label patients or low quality images are common in the test population, the performance of the algorithm will be lower than in the original dataset.[11] Finally, the study should report how the AI device was integrated into the trial setting, including how the results were interpreted or made available, and whether the interface and code can be accessed publicly.

When does it not work, and why?

This emphasis is perhaps more important for AI interventions than others, and arguably the most important issue raised by the AI extension guidelines. Evaluation of diagnostic accuracy metrics are not enough in isolation. AI interventions will rarely make the same mistakes as clinicians, so equal performance will not necessarily lead to equal outcomes. [8,12] Furthermore, AI interventions may often encode the biases of their human creators. The dangers of algorithmic harm are also potentially compounded by homogeneity and scale - if an AI system performs poorly on a certain disease and/or a certain population, this effect may be replicated around the world. By contrast, human decision makers might be biased but the effect may be mitigated, at least somewhat, by their diversity of biases.[13] These issues may be compounded by the fact that many AI algorithms are not interpretable. This wouldn't necessarily be a problem, as clinician's judgment is not always interpretable either, except that minor perturbations in input parameters can often unpredictably affect the output in a way that is non intuitive, and the causes of these errors are often not identified unless they are specifically looked for.[12] The recommendation is to "describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, explain why not." Since the variation in input parameters will almost always be higher in clinical practice than in a tightly regulated clinical trial, it is incumbent on the investigators to at least explore algorithm failures within the available data. This is an area of active translational work between computer scientists and clinicians as methods are developed both to train more robust networks that are less brittle with respect to input data, as well as more interpretable to improve the face validity of the results.

Although the CONSORT AND SPIRIT AI extensions are specifically attached to guidelines for reporting clinical trials (or protocols), it is worth noting that there are a number of parallel AI reporting guidelines that are currently underway on the diagnostic accuracy (e.g., STARD)[14] as well as risk prediction models (e.g., TRIPOD) (Table 1). [12,15] Given that all these reporting guidelines serve different purposes, it is important for the scientific and regulatory committee apply them appropriately on different stages of the AI development and testing. In summary, the standardization of reporting guidelines of AI will help 1) ensure validity, improve replicability, and maximize the utility of clinical research; 2) streamline and guide the approval pathway by the regulatory committee (e.g., US FDA, European CE and etc) [16] and; 3) improve patient safety, outcomes, and hopefully experience.[13] These guidelines are not substantively different from what we have established for medical devices or new drugs in the past, starting from phase 1 safety studies to phase 4 post-marketing surveillance studies. Taken together, these guidelines lay out a pragmatic pathway for rigorous evaluation not only of the efficacy of an algorithm but the effectiveness, equity, and safety of an AI device integrated into clinical care.

Table. Summary of guidelines for artificial intelligence studies. (Courtesy of Alistair Denniston and Xiaoxuan Liu)

Name of AI Extension	**Purpose of AI system** (e.g. diagnosis, prognosis, therapeutic decision-making, risk-stratification)	**Study design**	**Phase of development/testing of the AI system**	**Status**
Development and validation phase				
STARD-AI[14]	Diagnosis	Diagnostic accuracy study	Testing the diagnostic accuracy of an AI system	In development
TRIPOD-ML[15]	Diagnosis or prognosis	Studies developing, validating, or updating a prediction model	Development, validation and/or updating of an AI system	In development
Testing and regulatory phase				

CONSORT-AI[6]	Any health intervention	Randomized trial (report)	Randomized trial report, results for the effectiveness of an AI system	Published online September 9, 2020 in the British Medical Journal, Lancet Digital Health, and Nature Medicine.
SPIRIT-AI[7]	Any health intervention	Randomized trial (protocol)	Randomized trial protocol for testing the effectiveness of an AI system	Published online September 9, 2020 in the British Medical Journal, Lancet Digital Health, and Nature Medicine.

References

1. Center for Devices, Radiological Health. Artificial Intelligence and Machine Learning in Software. 2020. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [Accessed August 16, 2020].

2. Faes L, Liu X, Wagner SK, et al. A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Transl Vis Sci Technol 2020;9:7. Available at: http://dx.doi.org/10.1167/tvst.9.2.7.

3. Eaton LA. CONSORT Guidelines. Encyclopedia of Behavioral Medicine 2013:486–487. Available at: http://dx.doi.org/10.1007/978-1-4419-1005-9_638.

4. Moher D, Chan A-W. SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials). Guidelines for Reporting Health Research: A User's Manual 2014:56–67. Available at: http://dx.doi.org/10.1002/9781118715598.ch7.

5. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019;25:1467–1468. Available at: http://dx.doi.org/10.1038/s41591-019-0603-3.

6. Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston and the SPIRIT-AI and CONSORT-AI Working Group. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence. Nature Medicine, British Medical Journal, and Lancet Digital Health 2020.

7. Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert and the SPIRIT-AI and CONSORT-AI Working Group. Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension. Nature Medicine, British Medical Journal, and Lancet Digital Health 2020.

8. Kaplan RM, Irvin VL. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. PLoS One 2015;10:e0132382. Available at: http://dx.doi.org/10.1371/journal.pone.0132382.

9. Klonoff DC, Kerr D, Mulvaney SA. Diabetes Digital Health. Elsevier; 2020. Available at: https://play.google.com/store/books/details?id=p3bLDwAAQBAJ.

10. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683. Available at: http://dx.doi.org/10.1371/journal.pmed.1002683.

11. Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems 2020. Available at: http://dx.doi.org/10.1145/3313831.3376718.

12. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the ACM Conference on Health, Inference, and Learning 2020. Available at: http://dx.doi.org/10.1145/3368555.3384468.

13. Singh RP, Hom GL, Abramoff MD, et al. Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. Translational Vision Science & Technology 2020;9:45. Available at: http://dx.doi.org/10.1167/tvst.9.2.45.

14. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 2020;26:807–808. Available at: http://dx.doi.org/10.1038/s41591-020-0941-1.

15. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–1579. Available at: http://dx.doi.org/10.1016/S0140-6736(19)30037-6.

16. Harvey HB, Gowda V. How the FDA Regulates AI. Acad Radiol 2020;27:58–61. Available at: http://dx.doi.org/10.1016/j.acra.2019.09.017.