Published in final edited form as:

Ophthalmology. 2020 December ; 127(12): 1596–1599. doi:10.1016/j.ophtha.2020.09.009.

Reporting Guidelines for Artificial Intelligence in Medical Research

J. Peter Campbell, MD, MPH,
Portland, Oregon

Aaron Y. Lee, MD, MSci,
Seattle, Washington

Michael Abràmoff, MD, PhD,
Iowa City, Iowa

Pearse A. Keane, MD,
London, United Kingdom

Daniel S.W. Ting, MD, PhD,
Singapore

Flora Lum, MD,
San Francisco, California

Michael F. Chiang, MD
Portland, Oregon

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 because 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, may not report results similarly, or may not appreciate fully the potential for inaccurate conclusions based on interpretation errors. Thus, 2 key motivations exist for developing reporting standards for academic research involving novel technologies. First, nonstandard reporting may limit the validity, comparability, and usefulness of research; standardization improves the return on investment of all research efforts. Second, clinical decisions based on unfamiliar technology may cause harm, in the form of either 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 may arrive at conclusions that are not valid. In the second, we may arrive at 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 arrive at incorrect conclusions as a result. As clinicians, because of our obligation to *primum non nocere* and to ensure the bioethical principles of non-maleficence and justice, it is

Correspondence: Flora Lum, MD, American Academy of Ophthalmology, 655 Beach Street, San Francisco, CA 94109. flum@aao.org.

incumbent on us to understand emerging technologies as they relate to our clinical care for patients.

We currently are in the midst of an abundance of articles involving artificial intelligence (AI) algorithms in clinical medicine. Ophthalmology, with frequent use of ophthalmic imaging, has been on the forefront of this technological advancement. Between 2015 and 2020, 728 publications appeared that used the terms *artificial intelligence* or *deep learning*, with the ophthalmology discipline having 10 times as many in 2020 compared with 2017. As the United States Food and Drug Administration develops new pathways for regulatory approval,[1] appreciation is growing not only for the potential benefits of AI in clinical medicine, but also for the ways that it can fail, cause harm, or both when implemented.[2] It is important to note that although automated image-based diagnosis is a frequent application of AI in ophthalmology, the potential applications are myriad, including many other types of structured data. To facilitate the development of clinical AI devices that not only are effective in a research study, but also safe, effective, equitable, and reliable in practice, a relatively urgent need exists to standardize the reporting of AI studies by ensuring minimum necessary details for critical review, interpretation, and application of AI.

Originally designed for randomized clinical trials, the Consolidated Standards of Reporting Trials (CONSORT) and the accompanying Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) guidelines were developed to standardize reporting of clinical trials and clinical trial protocols, respectively.[3,4] They have been adopted widely by medical journals, streamlining and standardizing the review process, enhancing the ability to compare between trials, and improving the interpretability of clinical trial results overall. Several extensions to the original guidelines have been devised to address study designs that are not randomized clinical trials (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, an international effort has been made to develop AI extensions both to CONSORT and SPIRIT guidelines that are being published simultaneously in *Nature Medicine*, the *British Medical Journal*, and *Lancet Digital Health*.[6,7] Using a definition of AI as “the science of developing computer systems which can perform tasks normally requiring human intelligence,” the articles by Liu et al[6] and Rivera et al[7] meticulously describe the process for developing AI-specific guidelines that are considered essential for reporting AI clinical trials. These are meant to supplement the existing CONSORT and SPIRIT guidelines with 14 and 15 additional AI-specific recommendations, respectively. Until now, no requirement has been made for preregistration, such as on clinicaltrials.gov, although this has been shown to increase replication and to lower the effect size of studies compared with post hoc inclusion and exclusion as well as statistical analysis.[8] These AI-specific guidelines fall into 3 general categories or concepts that are important to understand and are discussed below.

What Is the Device and What Is It Intended to Do?

Several specific recommendations fall into this general category. First, to ensure transparency, the guidelines recommend specifying that the intervention involves AI within the title, abstract, or both. Second, the methods need to specify exactly what was studied

(hardware, software, version(s), etc.), including internal thresholds. Third, the indication for use needs to be defined explicitly, including by whom (e.g., who is the user) and where within the clinical pathway (e.g., as an autonomous device for disease screening, as an assistive diagnostic device for clinicians, to gauge prognosis after clinical diagnosis, etc). For example, although published AI algorithms can both (1) detect referable diabetic retinopathy and (2) specify the level of retinopathy, these are separately evaluable indications for use because each may be used 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 with 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 defined strictly 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 indication for use, and its integration into the clinical care pathway should be 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 3 clinical trial may not generalize to a population of patients who are 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) was studied. Specifically, the process for assessing and handling low-quality data needs to be defined. In addition, the methods should specify whether any human interaction was involved in selecting which inputs were studied and which were excluded. For example, in a real-world evaluation of an AI device for diabetic retinopathy screening, factors such as whether the technician turned off the lights in the room and waited a few seconds for appropriate pupillary dilation affected the outcome of the device.[11] In practical, real-world AI validation, these issues are critical, because many research datasets used for training are culled from low-quality images. These images may not perfectly demonstrate a class, label, or both. 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 is arguably the most important issue raised by the AI extension guidelines. Evaluation of diagnostic accuracy metrics is not enough in isolation. Artificial intelligence interventions rarely will make the same mistakes as clinicians, so equal performance will not necessarily lead to

equal outcomes.[8,12] Furthermore, AI interventions often may encode the biases of their human creators. The dangers of algorithmic harm also are compounded potentially by homogeneity and scale, that is, if an AI system performs poorly on a certain disease, a certain population, or both, this effect may be replicated around the world. By contrast, human decision makers may be biased, but the effect may be mitigated, at least somewhat, by their diversity of biases.[13] For example, fundus pigmentation varies significantly across ethnic groups. If this is not considered within training, 2 possible errors can occur. First, the algorithm may associate level of pigmentation incorrectly with the presence of disease if those are associated in the training population. Second, if the training data were not sufficiently heterogenous to the level of pigmentation in the fundus, the performance in populations that differ in this variable may decrease.[14] These issues may be compounded by the fact that many AI algorithms are not interpretable. This would not necessarily be a problem, because a clinician's judgment is not always interpretable either, except that minor perturbations in input parameters often can affect the output unpredictably and in a way that is nonintuitive, and the causes of these errors often are 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." Because the variation in input parameters almost always will be higher in clinical practice than in a tightly regulated clinical trial, it is incumbent on the investigators at least to explore algorithm failures within the available data. This is an area of active translational work between computer scientists and clinicians as methods are developed to train more robust networks that are less brittle with respect to input data and are more interpretable to improve the face validity of the results.

Although the CONSORT-AI and SPIRIT-AI extensions are specificially attached to guidelines for reporting clinical trials (or protocols), it is worth noting that the formation of a number of parallel AI reporting guidelines is currently underway on diagnostic accuracy (e.g., Standards for Reporting of Diagnostic Accuracy [STARD])[15] as well as risk prediction models (e.g., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis [TRIPOD]; Table 1).[12,16] Looking at the big picture, the standardization of reporting guidelines of AI will help (1) to ensure validity, improve replicability, and maximize the usefulness of clinical research; (2) to streamline and guide the approval pathway by the regulatory committee (e.g., United States Food and Drug Administration, European Conformité Européenne [CE], etc.)[17]; and (3) to improve patient safety, outcomes, and—we hope—experience.[13] Although the CONSORT-AI and SPIRIT-AI extensions are designed for those writing (and reviewing) manuscripts involving AI algorithms, we believe they are equally and perhaps more importantly relevant for those reading these manuscripts. 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 and ending with phase 4 post-marketing surveillance studies (also focused on patient safety). These guidelines lay out a pragmatic pathway for rigorous evaluation not only of the efficacy of an algorithm, but also the effectiveness, equity, and safety of an AI device integrated into clinical care.

Acknowledgments

Footnotes and Financial Disclosures

Financial Disclosure(s): The author(s) have made the following disclosure(s): J.P.C.: Financial support — Genentech.

A.Y.L.: Grants — Microsoft, NVIDIA, Carl Zeiss Meditec, Novartis, Genentech, Santen; Personal fees — Genetech, United States Food and Drug Administration, Verana Health, Santen.

M.A.: Consultant and Equity owner — Digital Diagnostics.

P.A.K.: Consultant — DeepMind Technologies, Inc, Roche, Novartis, Apellis; Financial support — Bayer; Lecturer — Bayer, Allergan, Topcon, Heidelberg Engineering, UK Research & Innovation.

D.S.W.T.: Consultant — Novartis; Steering committee — STARD-AI.

F.L.: Employee — American Academy of Ophthalmology.

M.F.C.: Consultant — Novartis; Grant — Genentech; Equity owner — InTeleretina, LLC.

Supported by the National Institutes of Health, Bethesda, Maryland (grant nos.: R01EY19474, R01EY031331, K12EY027720, and P30EY10572 [J.P.C. and M.F.C.]; K23EY029246 and NIH P30EY10572 [A.Y.L.]; and P30 EY025580 [M.A.]); Research to Prevent Blindness, Inc., New York, New York (unrestricted departmental funding [J.P.C., A.Y.L., M.A.] and Career Development Award [J.P.C.]); the National Science Foundation, Alexandria, Virginia (grant nos.: SCH-162279 [J.P.C. and M.F.C.]); the University of Iowa Hospitals and Clinics, Iowa City, Iowa (M.A.); and the UK Research & Innovation (Future Leaders Fellowship [P.A.K.]). The sponsors/ funding organizations had no role in the design or conduct of this research.

References

1. Center for Devices, Radiological Health, Food and Drug Administration. Artificial intelligence and machine learning in software. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device; 2020 Accessed 04.09.20.
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.
3. Eaton LA. CONSORT guidelines In: Gellman MD, Turner JR, eds. Encyclopedia of Behavioral Medicine. New York: Springer; 2013:486–487.
4. Moher D, Chan AW. SPIRIT (standard protocol items: recommendations for interventional trials) In: Moher D, Altman DG, Schulz KF, et al., eds. Guidelines for Reporting Health Research: A User's Manual. Hoboken, NJ: Wiley; 2014:56–67.
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. [PubMed: 31551578]
6. Cruz Rivera S, Liu X, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26: 1364–1374. [PubMed: 32908283]
7. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;370:m3210.
8. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. PLoS One. 2015;10:e0132382. [PubMed: 26244868]
9. Klonoff DC, Kerr D, Mulvaney SA. Diabetes Digital Health. Philadelphia: Elsevier; 2020.
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. [PubMed: 30399157]
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 10.1145/3313831.3376718; 2020 Accessed 04.09.20.

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 10.1145/3368555.3384468; 2020 Accessed 04.09.20.

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. Transl Vis Sci Technol. 2020;9:45.

14. Lee A. Machine diagnosis. Nature. 2019 Available at: https://www.nature.com/articles/d41586-019-01112-x.

15. 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. [PubMed: 32514173]

16. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577–1579. [PubMed: 31007185]

17. Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol. 2020;27:58–61. [PubMed: 31818387]

We currently are in the midst of an abundance of articles involving artificial intelligence algorithms in clinical medicine.

Table 1.

Summary of Guidelines for Artificial Intelligence Studies

Name of Artificial Intelligence Extension	Purpose of Artificial Intelligence System*	Study Design	Phase of Development or Testing of the Artificial Intelligence System	Status
Development and validation phase				
STARD AI[14]	Diagnosis	Diagnostic accuracy	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, or updating of an AI system, or a combination thereof	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*

AI = artificial intelligence; CONSORT = Consolidated Standards of Reporting Trials; ML = machine learning; SPIRIT = Standard Protocol Items: Recommendations for Interventional Trials; STARD = Standards for Reporting of Diagnostic Accuracy; TRIPOD = transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

Table courtesy of Alastair Denniston and Xiaoxuan Liu.

* For example, diagnosis, prognosis, therapeutic decision making, or risk stratification.