Journal of Dental Research
2021, Vol. 100(7) 677–680

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0022034521998337
journals.sagepub.com/home/jdr

Better Reporting of Studies on Artificial Intelligence: CONSORT-AI and Beyond

F. Schwendicke[1] and J. Krois[1]

Abstract

An increasing number of studies on artificial intelligence (AI) are published in the dental and oral sciences. The reporting, but also further aspects of these studies, suffer from a range of limitations. Standards towards reporting, like the recently published Consolidated Standards of Reporting Trials (CONSORT)-AI extension can help to improve studies in this emerging field, and the *Journal of Dental Research (JDR)* encourages authors, reviewers, and readers to adhere to these standards. Notably, though, a wide range of aspects beyond reporting, located along various steps of the AI lifecycle, should be considered when conceiving, conducting, reporting, or evaluating studies on AI in dentistry.

Keywords: clinical studies/trials, computer vision, decision-making, deep learning, personalized medicine, software engineering

Introduction

Artificial intelligence (AI) in healthcare receives an increasing amount of attention, mainly as AI-based applications are considered to have the potential for making care better, more efficient, and affordable, and hence, accessible. However, the quality of "AI for health" studies remains low, and reporting is often insufficient to fully comprehend and possibly replicate these studies (Liu et al. 2019; Nagendran et al. 2020; Wynants et al. 2020). Given the emergence of studies on AI in dentistry, action seems needed (Schwendicke et al. 2019).

CONSORT-AI

For randomized controlled trials (RCTs) and their protocols, reporting standards have been introduced and updated over the last 25 y. The CONSORT (Consolidated Standards of Reporting Trials) statement provides evidence-based recommendations for reporting of RCTs, while the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) statement guides reporting of RCT protocols (Moher et al. 2010, 2015). Both have been adopted by the vast majority of journals as expected standards and adhering to CONSORT has been found to increase reporting quality (Plint et al. 2006). Since RCTs are widely used to inform decision-makers in health policy, regulations, and clinical care, their comprehensive and systematic reporting is crucial, as only can then stakeholders gauge a trial's methodology, validity, and bias, or attempt replication.

For AI studies in healthcare, only a limited number of RCTs are available (Kelly et al. 2019). Notably, such RCTs have reported mixed results on the efficacy and applicability of AI, often in contrast to more optimistic retrospective studies assessing AI applications (for details, see references in Liu et al. (2020)). A rigorous and prospective evaluation of AI interventions should hence be expected—in the same way we expect other medical interventions to be evidence-based, with proven benefit and safety.

For reporting RCTs and RCT protocols involving AI, new extensions of the SPIRIT and CONSORT guidelines have been recently published (Liu et al. 2020; Rivera et al. 2020). The *Journal of Dental Research (JDR)* encourages authors to consult these guidelines and employ them to comprehensively and transparently lay out the planned or concluded trial methodology and findings, and urges reviewers to check any AI submissions in the *JDR* against these guidelines.

Beyond Reporting of Randomized Trials

However, better reporting of RCTs is not sufficient; a range of aspects over the lifecycle of AI for health interventions should be considered (Fig. 1).

1. The majority of studies in AI in dentistry are not RCTs. They should nevertheless be conducted and reported at the highest standards. In the absence of existing standards, authors, reviewers, and readers of the *JDR* should consult other guidelines like the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) (Mongan

[1] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin Berlin, Berlin, Germany

A supplemental appendix to this article is available online.

Corresponding Author:
F. Schwendicke, Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin Berlin, Aßmannshauser Str. 4-6, Berlin 14197, Germany.
Email: falk.schwendicke@charite.de

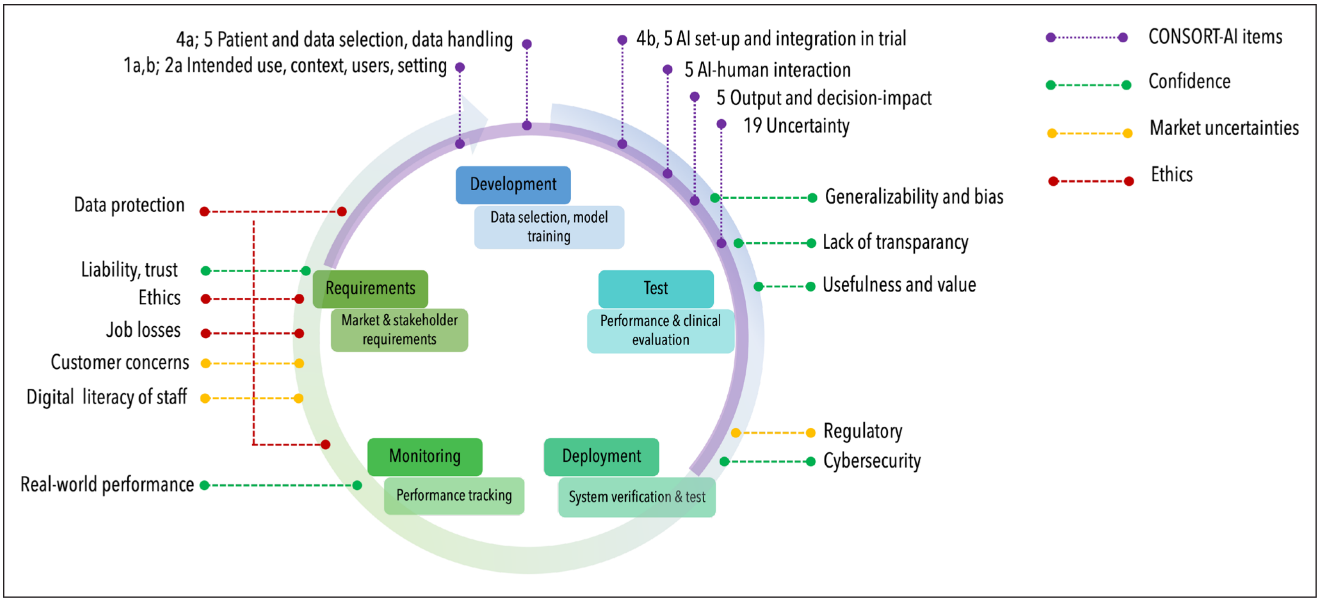
4a; 5 Patient and data selection, data handling
1a,b; 2a Intended use, context, users, setting
4b, 5 AI set-up and integration in trial
5 AI-human interaction
5 Output and decision-impact
19 Uncertainty
CONSORT-AI items
Confidence
Market uncertainties
Ethics
Data protection
Liability, trust
Ethics
Job losses
Customer concerns
Digital literacy of staff
Real-world performance
Development
Data selection, model training
Requirements
Market & stakeholder requirements
Test
Performance & clinical evaluation
Monitoring
Performance tracking
Deployment
System verification & test
Generalizability and bias
Lack of transparancy
Usefulness and value
Regulatory
Cybersecurity

Figure. The lifecycle of artificial intelligence (AI) applications usually begins with an assessment of requirements, followed by development, testing, deployment of the AI, to monitoring in clinical care and reassessment. Different aspects along this lifecycle, acting as barriers to adoption of AI applications, have been identified (Ammanath et al. 2020); confidence into AI, market uncertainties, ethical concerns. These are located along the lifecycle (indicated via green, yellow, and red dotted lines). The Consolidated Standards of Reporting Trials (CONSORT)-AI extension items (purple dotted lines, e.g., items 1a,b; 2a; 4a,b; 5; 19) cover mainly the development and test steps, specifically when reporting a randomized controlled trial (purple semicircle).

et al. 2020) or the STARD (Standards for Reporting of Diagnostic Accuracy Studies) (Bossuyt et al. 2015) statements. We will soon see statements targeting other study types (like modelling studies) with a focus on AI (Collins and Moons 2019).

2. Better reporting does not mitigate poor design and conduct. One weakness encountered in AI for health research is the so-called "AI chasm" (Keane and Topol 2018), with accuracies not necessarily relating to patients' or health systems' benefit. Researchers should aim to overcome this by reporting accuracy comprehensively, including sensitivity, specificity, true and false positive rates, positive and predictive values, and receiver-operating characteristics curves as well as metrics robust against imbalances, like the F1-score (Krois et al. 2019) or the area under the precision-recall curve (Saito and Rehmsmeier 2015). Notably, the appropriate set of model performance metrics depends on the formulation of the machine learning task itself. For instance, in the field of computer vision object detection or segmentation models are more appropriately evaluated by the Intersection-over-the-Union (IoU) or mean average precision (mAP). In particular, for pixel level (segmentation) models the translation into toothwise metrics should be attempted to ease interpretation by medical professionals (Cantu et al. 2020). Also, accuracies should be translated into tangible value (health benefit, costs); the long-term effects of AI should be explored, for example using model-based extrapolations (Schwendicke et al. 2020). The impact of AI on the clinical workflow, on decision-making as well as its acceptability, fidelity, and maintenance should be considered. Gauging the relationship between the user or receiver of AI (dentists, patients) and the AI intervention in a clinical environment is required (Kelly et al. 2019). Researchers may want to adopt the perspective on many AI applications as "complex interventions" rather than narrowly defined diagnostic or predictive tools, and consider validated and accepted frameworks accordingly (Moore et al. 2015).
3. Any kind of results should be explored for their generalizability; the "transportability" of AI to different populations (disease manifestations, prevalence) or data sources (e.g., electronic health records, radiographic machinery) should be demonstrated before considering translation into clinical studies or even care. Validation on external data that represent the breadth of potential target populations is required, ideally demonstrating temporal, spectral, and geographic transportability (König et al. 2007).
4. Model bias may have a vast range of sources, for example representation, data-snooping, measurement, label, spectrum, evaluation, or deployment bias. A range of methods and risk of bias tools for assessing diagnostic accuracy and prediction studies like QUADAS-2 (Whiting et al. 2011) and PROBAST (Wolff et al. 2019) can be employed to gauge bias. Bias can further be detected via stratification, subgroup and clustering analyses, or regression analyses, and associated methods (Adebayo 2016; Badgeley et al. 2019).

Moreover, methods to compare the similarity of the training dataset and real-world data (thereby determining representativeness) are available (Campagner et al. 2020). For dealing with detected bias, researchers may employ re-, under- and over-sampling techniques. Open-source toolkits may be employed to detect and mitigate bias (Bellamy et al. 2019). It is conceivable that in the future AI-based systems for healthcare will be audited by authoritative bodies for bias, interpretability, robustness, and possible failure modes, among others (Oala et al. 2020). Providing scripts, code, and datasets (open research) for reproducing the results should be considered wherever possible under data protection and intellectual property considerations. However, the complexity of AI studies and the interdependencies of code, data, and the computational environment may impede a reanalysis thereof. Hence, it is more likely that the usage of platforms for benchmarking, including the assessment of bias and trustworthiness of AI models will be considered a mandatory, alternative model validation step.

5. Given that many AI applications involve highly complex prediction models which are hard to interpret, researchers should aim to include elements of explainability (XAI) and transparency into their studies (Lapuschkin et al. 2019). XAI enables to detect and counteract bias, allows liability and accountability and supports fairness, generalizability, and transportability.
6. Lastly, clearly communicating the different impacts of AI onto clinical care allows all stakeholders to make informed decision towards the adoption of AI. Model fact labels (comparable to food labeling) are examples supporting such communication (Sendak et al. 2020).

Standards like the recently published CONSORT-AI extensions will improve reporting of AI studies in dentistry, and the *Journal of Dental Research* encourages authors, reviewers, and readers to adhere to these standards. A range of further aspects along the AI lifecycle, laid out above, should be considered when conceiving, conducting, reporting, or evaluating studies on AI in dentistry.

Author Contributions

F. Schwendicke, contributed to conception, design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript; J. Krois, contributed to data acquisition, analysis, and interpretation, critically revised the manuscript. Both authors gave final approval and agree to be accountable for all aspects of the work.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors are co-founders of a startup on dental image analysis using AI. The conception and writing of this article were independent from this.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

F. Schwendicke https://orcid.org/0000-0003-1223-1669

References

Adebayo J. 2016. FairML: toolbox for diagnosing bias in predictive modeling. Massachusetts Institute of Technology [accessed 2021 Aug 2]. https://dspace.mit.edu/handle/1721.1/108212?show=full.

Ammanath B, Novak DR, Anderson S, Kulkarni A. 2020. Conquering AI risks. Deloitte Insights [accessed 2021 Aug 2]. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/conquering-ai-risks.html.

Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2:31.

Bellamy RKE, Dey K, Hind M, Hoffman S, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilović A, et al. 2019. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev [epub ahead of print 7 Nov 2020]. doi:10.1147/JRD.2019.2942287

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, et al. 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 351:h5527.

Campagner A, Sconfienza L, Cabitza F. 2020. H-accuracy, an alternative metric to assess classification models in medicine. Stud Health Technol Inform. 270:242–246.

Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F. 2020. Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent. 100:103425.

Collins GS, Moons KGM. 2019. Reporting of artificial intelligence prediction models. Lancet. 393(10181):1577–1579.

Keane PA, Topol EJ. 2018. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 1:40.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17(1):195.

König IR, Malley JD, Weimar C, Diener HC, Ziegler A. 2007. Practical experiences on the necessity of external validation. Stat Med. 26(30):5499–5511.

Krois J, Graetz C, Holtfreter B, Brinkmann P, Kocher T, Schwendicke F. 2019. Evaluating modeling and validation strategies for tooth loss. J Dent Res. 98(10):1088–1095.

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. 2019. Unmasking clever hans predictors and assessing what machines really learn. Nat Commun. 10(1):1096.

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, et al. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 1(6):e271–e297.

Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Chan A-W, Darzi A, Holmes C, Yau C, Ashrafian H, et al. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 26(9):1364–1374.

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 340:c869.

Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA; PRISMA-P Group. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 4(1):1.

Mongan J, Moy L, Kahn CE Jr. 2020. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology. 2(2):e200029.

Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, Moore L, O'Cathain A, Tinati T, Wight D, et al. 2015. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 350:h1258.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 368:m689.

Oala L, Fehr L, Gilli L, Balachandran P, Leite AW, Calderon-Ramirez S, Li DX, Nobis G, Muñoz A, Jaramillo-Gutierrez G, et al. 2020. ML4H auditing: from paper to practice. Paper presented at: Proceedings of the Machine Learning for Health NeurIPS Workshop; 2020 Dec 11-12; online. 136:280–317. https://neurips.cc/virtual/2020/protected/workshop_16134.html

Plint AC, Moher D, Morrison A. 2006. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust. 185(5):263–267.

Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ. 370:m3210.

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 10(3):e0118432.

Schwendicke F, Golla T, Dreher M, Krois J. 2019. Convolutional neural networks for dental image diagnostics: a scoping review. J Dent. 91:103226.

Schwendicke F, Rossi JG, Göstemeyer G, Elhennawy K, Cantu AG, Gaudin R, Chaurasia A, Gehrung S, Krois K. 2020. Cost-effectiveness of artificial intelligence for proximal caries detection. J Dent Res [epub ahead of print 16 Nov 2020]. doi:10.1177/0022034520972335

Sendak MP, Gao M, Brajer N, Balu S. 2020. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 3:41.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 155(8):529–536.

Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 170(1):51–58.

Wynants L, Smits LJM, Van Calster B. 2020. Demystifying AI in healthcare. BMJ. 370:m3505.