Guidelines for clinical trials using artificial intelligence – SPIRIT-AI and CONSORT-AI†

Clare McGenity1,2* and Darren Treanor1,2,3,4

1 Leeds Teaching Hospitals NHS Trust, Leeds, UK
2 University of Leeds, Leeds, UK
3 Department of Clinical Pathology and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
4 Centre for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden

*Correspondence to: C McGenity, Department of Histopathology, St. James’ University Hospital, Beckett Street, Leeds LS9 7TF, UK
E-mail: clare.mcgenity@nhs.net

†Invited commentary for Cruz Rivera et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence. The SPIRIT-AI extension. Nat Med 2020; 26: 1351–1363 and Liu et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence. The CONSORT-AI extension. Nat Med 2020; 26: 1364–1374.

Abstract

The rapidly growing use of artificial intelligence in pathology presents a challenge in terms of study reporting and methodology. The existing guidelines for the design (SPIRIT) and reporting (CONSORT) of clinical trials have been extended with the aim of ensuring production of the highest quality evidence in this field. We explore these new guidelines and their relevance and application to pathology as a specialty.

© 2020 The Authors. The Journal of Pathology published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland.

Keywords: artificial intelligence; CONSORT-AI; SPIRIT-AI; digital pathology; pathology; clinical trial; randomised trial; reporting guidelines; checklist; machine learning

Received 28 August 2020; Accepted 29 September 2020

No conflicts of interest were declared.

Although the word ‘revolution’ is somewhat overused in technology circles, the recent leap in performance of artificial intelligence (AI) systems surely does justify the term. Driven by advances in a particular type of neural network called ‘deep learning’ [1], computers have achieved human-level performance in a number of tasks previously considered to be some decades in the future [1–3].

Relevance to pathology

The area of pathological diagnosis has been included in this revolution [4] and arguably pathology data (and specifically image interpretation) are ideally suited to the application of deep learning, which at its core is a pattern-recognition tool ‘trained’ on data to classify new ‘test’ data. In a short period of time, we have seen the technology applied successfully in a variety of applications, with resulting histopathology-focused papers in high impact general medical and science journals [5–11], many claiming pathologist-level performance.

But AI is neither magical nor truly ‘intelligent’ like a human. Despite impressive results in test datasets under controlled conditions, in real-world applications it does not always deliver according to the hype and excitement of initial discoveries. This ‘brittleness’ has a variety of causes, including over-sensitivity to training data, lack of variety and depth in training sets, and failure to anticipate real-world conditions of deployment [12,13]. Many studies to date have been small, remote from real-world clinical use, and actual real-world application of AI in pathology is exceptionally rare.

The consequences of this are serious – a possible ‘replication crisis’ in digital pathology AI, and worse still, clinical harm due to the use of inaccurate or unreliable AI systems in clinical practice without proper oversight. The novelty of AI and relative inexperience of our community with the technology combines with the commercial pressure on AI companies to show positive results and the publication pressures on academic pathologists to create a potentially serious risk.

New guidelines recently published will go some way to alleviate this risk. The EQUATOR network was founded to bring together researchers, medical journal editors, peer reviewers, developers of reporting guidelines, research funding bodies and other collaborators with mutual interest in improving the quality of research.
publications and of research itself [14]. The EQUATOR mission is to achieve accurate, complete and transparent reporting of all health research studies to support research reproducibility and usefulness [14,15]. To address potential issues around AI, extensions to the SPIRIT and CONSORT guidelines were registered as ‘guidelines under development’ with the EQUATOR network in 2019 [16,17].

**SPIRIT–AI and CONSORT–AI guidelines**

Using a systematic approach with domain experts and methodologists, the existing guidelines for the design (SPIRIT) and reporting (CONSORT) of clinical trials have been modified to address the challenges provided by AI. The guidelines have been extended to include 15 and 14 new items, respectively, covering areas such as:

- The need to clearly describe the intended use of the AI intervention
- Indicators for how to use the AI intervention in the clinical setting
- Details on the data inputs to train the AI tool, and the outputs it produces
- Descriptions of how errors or failures of the system are reviewed
- Human–computer interaction aspects of the AI intervention

The intention of the guidelines is not to be prescriptive or reduce innovation, but to improve the consistency of the design and reporting of research in this area and improve transparency so that systems and results can be more easily evaluated. As such, the guidelines offer a much-needed framework in which researchers can frame their plans to evaluate AI technologies, which will drive up the quality of research in this area. The authors acknowledge that this is a rapidly evolving area and there will probably need to be frequent reviews and updates of the guidelines.

There are several areas for future work – despite the publicity around AI, only seven clinical trials of AI have published results on clinicaltrials.gov (that is across all domains, and none in histopathology [17]). So, as evidence and experience accumulate, trial design and reporting will probably become more sophisticated. Relatively little work has been carried out using AI in pathology and more domain-specific recommendations may be needed. Finally, the guidelines specifically exclude the reporting of ‘continuously improving’ AI, as this is a more novel method that may require a different (revolutionary!) approach to design and reporting.

**Conclusions**

As we sit at the precipice of a technological transformation in the use of AI within pathological assessment and diagnosis, a quote from Alan Turing (considered the father of modern computing and AI) in *The Times* newspaper of 11th June 1949 remains pertinent: ‘This is only a foretaste of what is to come, and only the shadow of what is going to be’. Nonetheless, in the urgency to develop these technologies, we must at the same time recall our Hippocratic Oath to ‘do no harm’ and ensure we create the best quality evidence for the benefit of our patients.

**Acknowledgements**

We thank Dr Xiaoxuan Liu and Dr Alastair Denniston for their advice and proof reading.

Dr McGenity is funded by Leeds Cares (https://leedscares.org/). Dr Treanor is funded by the National Pathology Imaging Co-operative (NPIC) (https://npic.ac.uk/). NPIC (project no. 104687) is supported by the Data to Early Diagnosis and Precision Medicine strand of the UK Government’s Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI).

**Author contributions statement**

CM and DT designed, drafted and edited this document together.

**References**

1. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019; 7: 53040–53065.
2. Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks and tree search. *Nature* 2016; 529: 484–489.
3. Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning. *arXiv: 1312.5602 [cs:LG]*; December 2013.
4. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence – the third revolution in pathology. *Histopathology* 2019; 74: 372–376.
5. Bejnordi BE, Veta M, van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; 318: 2199–2210.
6. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019; 2: 48.
7. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; 25: 1054–1056.
8. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; 25: 1301–1309.
9. Bulten W, Pinckaers H, van Boven H, et al. Automated Gleason grading of prostate biopsies using deep learning. *Lancet Oncol* 2020; 21: 233–241.
10. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* 2020; 1: 800–810.
11. Strom P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based diagnostic study. *Lancet Oncol* 2020; 21: 222–232.
12. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med* 2019; 2: 25.

13. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; 17: 195.

14. The EQUATOR Network and UK EQUATOR Centre. EQUATOR Network. [Accessed 24 August 2020]. Available from: https://www.equator-network.org

15. Simera I, Altman DG, Moher D, et al. Guidelines for reporting health research: the EQUATOR network’s survey of guideline authors. *PLoS Med* 2008; 5: e139.

16. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence. The SPIRIT-AI extension. *Nat Med* 2020; 26: 1351–1363.

17. Liu X, Cruz Rivera S, 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.