A Quality Assessment Tool for Artificial Intelligence Centred Diagnostic Test Accuracy Studies: QUADAS-AI

Viknesh Sounderajah[1,2], Hutan Ashrafian[1,2], Sherri Rose[7], Nigam H. Shah[34], Marzyeh Ghassemi[3], Robert Golub[6], Charles E. Kahn, Jr.[19], Andre Esteva[17], Alan Karthikesalingam[8], Bilal Mateen[18], Dale Webster[8], Dan Milea[20], Daniel Ting[20], Darren Treanor[21,22,23,24], Dominic Cushnan[25], Dominic King[1,26], Duncan McPherson[27], Ben Glocker[36], Felix Greaves[28], Leanne Harling[1,2], Johan Ordish[27], Jérémie F. Cohen[29], Jon Deeks[5], Mariska Leeflang[13], Matthew Diamond[30], Matthew D.F. McInnes[31], Melissa McCradden[32], Michael D. Abràmoff[33], Pasha Normahani[2], Sheraz R. Markar[2], Stephanie Chang[35], Xiaoxuan Liu [14,15,16], Susan Mallett[4], Shravya Shetty[8], Alastair Denniston[14,15,16], Gary S. Collins[9,10], David Moher[11], Penny Whiting[12], Patrick M. Bossuyt*[13] and Ara Darzi*[1,2]

* Joint senior authorship

Author Affiliations:

[1] Institute of Global Health Innovation, Imperial College London, United Kingdom
[2] Department of Surgery and Cancer, Imperial College London, United Kingdom
[3] Institute for Medical Engineering & Science, Massachusetts Institute of Technology, United States of America
[4] Centre for Medical Imaging, University College London, United Kingdom
[5] Institute of Applied Health Research, University of Birmingham, United Kingdom
[6] JAMA (Journal of the American Medical Association), United States of America
[7] Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, California, United States of America
[8] Google Health
[9] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom
[10] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
[11] Ottawa Hospital Research Institute, Canada
[12] Bristol Medical School, University of Bristol, Bristol, United Kingdom
[13] Department of Epidemiology and Data Science, Amsterdam University Medical Centres, University of Amsterdam, The Netherlands
[14] Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, United Kingdom
[15] University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
[16] Health Data Research UK, London, United Kingdom
[17] Salesforce Research, San Francisco, United States of America
[18] Wellcome Trust, London, United Kingdom
[19] University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
[20] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[21] Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom
[22] University of Leeds, Leeds, UK
[23] Department of Clinical Pathology, and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
[24] Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden
[25] NHSX, London, United Kingdom
[26] Optum, Paddington, London, United Kingdom
[27] Medicines and Healthcare products Regulatory Agency, London, United Kingdom
[28] National Institute for Health and Care Excellence, London, United Kingdom

[29] Department of Pediatrics and Inserm UMR 1153 (Centre of Research in Epidemiology and Statistics), Necker - Enfants Malades Hospital, Assistance Publique - Hôpitaux de Paris, Université de Paris, Paris, France
[30] Food and Drug Administration, Maryland, United States of America
[31] Departments of Radiology and Epidemiology, University of Ottawa, The Ottawa Hospital Research Institute, Canada
[32] Department of Bioethics, The Hospital for Sick Kids, Toronto, Canada
[33] Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa, United States of America
[34] Center for Biomedical Informatics Research, Stanford University, California, United States of America
[35] Annals of Internal Medicine, American College of Physicians, United States of America
[36] Faculty of Engineering, Department of Computing, Imperial College London, United Kingdom

Corresponding authors
Ara Darzi.
Institute of Global Health Innovation & Department of Surgery and Cancer, Imperial College London, United Kingdom. a.darzi@imperial.ac.uk

Patrick M. Bossuyt
Department of Epidemiology and Data Science, Amsterdam University Medical Centres, University of Amsterdam, The Netherlands. p.m.bossuyt@amsterdamumc.nl

To the Editor - Over the next decade, it is projected that artificial intelligence (AI), particularly machine learning, centred systems will become key components of several workflows within the health sector. Medical diagnosis is particularly seen as one of the first areas that would harbour the adoption of AI innovations. Indeed, over 90% of health-related AI systems that have reached regulatory approval by the U.S. Food and Drug Administration belong to the field of diagnostics [1].

In the current paradigm, the majority of diagnostic investigations require interpretation from a clinician to identify the presence of a target condition; a crucial step in determining subsequent treatment strategies. Despite being an essential step in the provision of patient care, many health systems find it increasingly difficult to meet the demand for diagnostic test interpretation. To address this issue, diagnostic AI systems have been characterised as medical devices which may alleviate the burden placed upon diagnosticians: serving as case triage tools, enhancing diagnostic accuracy, and stepping in as a second reader when necessary. As AI centred diagnostic test accuracy (AI DTA) studies emerge, there has been a concurrent rise in systematic reviews which amalgamate the findings of comparable studies.

Strikingly, of these published AI DTA systematic reviews, 94% have been conducted in the absence of an AI specific quality assessment tool [2]. The most commonly used instrument is the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) tool [3]. QUADAS-2 is a risk of bias and applicability tool whose use is encouraged by PRISMA 2020 guidance [4]. QUADAS-2 does not, however, accommodate for niche terminology encountered in AI DTA studies nor does it signal researchers to the sources of bias found within this class of studies. Examples of such biases, when framed against the established domains of QUADAS-2 (Patient Selection; Index Test; Reference Standard; and Flow and Timing) are listed in table 1.

Domain	Description	Biases
Patient Selection	A description of included patients detailing prior	In AI DTA studies, eligible patients are often excluded on account of competing input data entry requirements (e.g. image quality) which, themselves, are variably reported. As highlighted by the CONSORT-AI

	testing, presentation, setting and the intended use of the index test.	guidelines [5], there is a need to accurately characterise the source, size and quality of input data alongside clear patient eligibility criteria. Data source issues can negatively impact the performance and overall applicability of an index test. For example, to minimise research costs, there has been increasing usage of datasets sourced from open-source repositories. Whilst this offers a pragmatic option, many open-source datasets have been found to house the inadvertent duplication of data across repositories, erroneous labelling, and incomplete patient demographic data. Manuscripts reporting both the development and validation of an index test rarely present the rationale and breakdown of its training, validation, and test sets. Small datasets, particularly those that lack complexity and balance, can result in overfitting, whereby the final index test resembles the training data too closely and is unable to reliably fit additional data. The clinical manifestation of this issue is the inability to accurately diagnose instances of a pathology if its clinical presentation does not closely resemble the training cases that the index test had previously encountered. There are various points within the data curation pipeline where quality may be compromised. For example, image pre-processing, a practice whereby image formats and resolutions are homogenised for the purpose of training, is an essential step in AI workflows. However, either down- or up-scaling resolution may impact the ability of certain index tests to identify diagnostic features effectively. Moreover, the lack of image metadata can also preclude the ability to explore an index test's dependence on specific data acquisition parameters, for example, the model of scanner used to acquire imaging data.
Index Test	The diagnostic test being evaluated and how it has been conducted and interpreted within the context of the study.	Only a limited number of published studies have undertaken adequate external evaluation when presenting the development and evaluation of their diagnostic tests. Reliance upon data from the same dataset that is used to train the diagnostic test (internal holdout set) can overestimate diagnostic performance.
Reference Standard	The choice of reference standard and how it has been conducted and interpreted within the context of the study.	There are multiple instances, as highlighted by Harris et al. [6], in which studies have reported the development of index tests against inappropriate reference standards, as opposed to more appropriate tests that provide higher sensitivity and specificity. For example, a clinician using a chest X-ray to diagnose pulmonary tuberculosis rather than the more accurate use of sputum culture. Studies with inappropriate reference standards are poorly reflective of real-world clinical practice in which reference standards consist of the amalgamation of clinical, radiological and laboratory data.

Flow and Timing	The time interval and the use of any interventions between the application of the index test and the reference standard	The timing between index test and reference standard is often poorly reported. As highlighted in a recent systematic review [7], studies which reported the performance of index tests to diagnose SARS-CoV-2 from chest X-rays did not routinely note the timing of the confirmatory RT-PCR test in relation to the imaging data. It is well understood that RT-PCR is a time sensitive assay and failing to report this relationship significantly hinders the overall clinical validity of the study results.

Table 1: Examples of bias within AI DTA studies

In order to tackle the sources of bias described above, as well as AI specific examples such as algorithmic bias, we propose an AI-specific extension to QUADAS-2 and QUADAS-C [8], a risk of bias tool developed for comparative accuracy studies. This new tool, QUADAS-AI, will provide researchers and policy makers with a specific framework to evaluate the risk of bias and applicability when conducting reviews evaluating AI DTA and reviews of comparative accuracy studies evaluating at least one AI centred index test.

QUADAS-AI will be complementary to ongoing reporting guideline tool initiatives, such as STARD-AI [9] and TRIPOD-AI [10]. QUADAS-AI is being coordinated by a global Project Team and Steering Committee consisting of clinician scientists, computer scientists, epidemiologists, statisticians, journal editors, EQUATOR Network representatives, regulatory leaders, industry leaders, funders, health policy makers and bioethicists. Given the reach of AI technologies, we view that connecting global stakeholders is of the utmost importance for this initiative. In turn, we would welcome contact from any new potential collaborators.

References

1. Benjamens, S., Dhunnoo, P. & Meskó, B. *npj Digit. Med.* **3**, 118 (2020).
2. Jayakumar, S. *et al. What are the Quality Assessment Standards used in Artificial Intelligence Diagnostic Accuracy Systematic Reviews?* https://www.researchsquare.com (2021) doi:10.21203/RS.3.RS-329433/V1.
3. Whiting, P. F. QUADAS-2: *Ann. Intern. Med.* **155**, 529 (2011).
4. Page, M. J. *et al.* The PRISMA 2020 statement: *The BMJ* vol. 372 (2021).
5. Liu, X. & Rivera, S. C. *Nat. Med. 2020 269* **26**, 1364–1374 (2020).
6. Harris, M. *et al. PLoS One* **14**, (2019).
7. Roberts, M. *et al. Nat. Mach. Intell.* **3**, 199–217 (2021).
8. Yang, B. *et al.* (2018) doi:10.17605/OSF.IO/HQ8MF.
9. Sounderajah, V. *et al. Nature Medicine* vol. 26 807–808 (2020).
10. Collins, G. & Moons, K. *Lancet* **393**, 1577–1579 (2019).

Acknowledgments

Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research Centre (BRC). GSC is supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294). DT is funded by National Pathology Imaging Co-operative, NPIC (Project no. 104687), supported by a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI). FG is supported by the National Institute for Health Research Applied Research Collaboration Northwest London. The views and opinions expressed herein are those of the authors and do not necessarily reflect the views of their employers or funders.

Author contributions

VS, SR, NS, MG, RG, CEK, XL, GSC, DW, AE, HA, DM (Dan Milea), DM (Duncan McPherson), JO, DT, JFC, ML, MM, MDFM, MDA, SM, PW and PMB prepared the first draft of the manuscript. Critical edits and feedback have been attained from all co-authors. The study described in the manuscript has been conceptualised, discussed, and agreed upon between all co-authors.

Competing interests

AK, SS and DW are employees at Google. AD and HA are employees at Flagship Pioneering UK Ltd. AE is an employee at Salesforce. DK is an employee at Optum. None of the other authors have any competing interests.