civic-health organizations, care professionals must help build a healthier future for all by empowering patients to have a voice in the policies affecting their wellbeing. The path to equity and improved health can be paved through physicians promoting their patients' right to vote. ❐

Simar Singh Bajaj[1], Alister Francois Martin[2,3] and Fatima Cody Stanford[4,5,6,7]✉

[1]*Department of the History of Science, Harvard University, Cambridge, MA, USA.* [2]*Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA.* [3]*Center for Social Justice and Health Equity, Boston, MA, USA.* [4]*MGH Weight Center, Massachusetts General Hospital, Boston, MA, USA.* [5]*Division of Endocrinology-Neuroendocrine, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.* [6]*Division of Endocrinology, Department of Pediatrics, Massachusetts General Hospital, Boston, MA, USA.* [7]*Nutrition Obesity Research Center at Harvard (NORCH), Boston, MA, USA.*

✉*e-mail: fstanford@mgh.harvard.edu*

Published online: 12 October 2021
https://doi.org/10.1038/s41591-021-01523-2

References

1. Voting Laws Roundup: July 2021. *Brennan Center for Justice* https://www.brennancenter.org/our-work/research-reports/voting-laws-roundup-july-2021 (2021).
2. Brown, C. L., Raza, D. & Pinto, A. D. *Public Health Rev.* **41**, 16 (2020).
3. Shelton, A. *Power Coalition for Equity and Justice* https://powercoalition.org/oped-voting-as-a-social-determinant-of-health/ (2021).
4. *Racial and Ethnic Health Disparities: What State Legislators Need to Know* (National Conference of State Legislators, 2013); https://www.ncsl.org/portals/1/documents/health/HealthDisparities1213.pdf
5. Greenwood-Ericksen, M. B. & Kocher, K. *JAMA Netw. Open* **2**, e191919 (2019).
6. Voting and Registration in the Election of November 2020. *The United States Census Bureau* https://www.census.gov/data/tables/time-series/demo/voting-and-registration/p20-585.html (2021).
7. Second Annual Civic Health Month Launches August 1. *Civic Health Month* https://www.civichealthmonth.org/press-releases (2021).
8. Martin, A., Raja, A. & Meese, H. *Int. J. Emerg. Med.* **14**, 26 (2021).
9. Liggett, A., Sharma, M., Nakamura, Y., Villar, R. & Selwyn, P. *Ann. Fam. Med.* **12**, 466–469 (2014).

Acknowledgements

F.C.S. received funding from National Institutes of Health and Massachusetts General Hospital Executive Committee on Research, National Institutes of Health NIDDK P30 DK040561 and L30 DK118710.

Author contributions

S.S.B. drafted the initial manuscript. A.F.M. and F.C.S. provided mentorship and editorial support. All authors read and approved the final draft before submission.

Competing interests

A.F.M. is the founder and executive director of VotER. The other authors declare no competing interests.

Check for updates

A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI

To the Editor — Over the next decade, systems that are centered on artificial intelligence (AI), particularly machine learning, are predicted to become key components of several workflows within the health sector. Medical diagnosis is seen as one of the first areas that may be revolutionized by AI innovations. Indeed, more than 90% of health-related AI systems that have reached regulatory approval by the US Food and Drug Administration belong to the field of diagnostics[1].

In the current paradigm, most 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 the interpretation of diagnostic tests. To address this issue, diagnostic AI systems have been characterized as medical devices that may alleviate the burden placed on diagnosticians: by serving as case triage tools, enhancing diagnostic accuracy and stepping in as a second reader when necessary. As AI-centered diagnostic test accuracy (AI DTA) studies emerge, there has been a concurrent rise in systematic reviews that amalgamate the findings of comparable studies.

Notably, 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 quality assessment of diagnostic accuracy studies (QUADAS-2) tool[3]. QUADAS-2 is a tool that assesses bias and applicability and its use is encouraged by current PRISMA 2020 guidance[4]. However, QUADAS-2 does not accommodate for niche terminology encountered in AI DTA studies, nor does it signal researchers to the sources of bias found within this class of study. 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.

To tackle these sources of bias, as well as AI-specific examples such as algorithmic bias, we propose an AI-specific extension to QUADAS-2 and QUADAS-C[5], a risk of bias tool that has been developed for comparative accuracy studies. This new tool, termed QUADAS-AI, will provide researchers and policy-makers with a specific framework to evaluate the risk of bias and applicability when conducting reviews that evaluate AI DTA and reviews of comparative accuracy studies that evaluate at least one AI-centered index test.

QUADAS-AI will be complementary to ongoing reporting guideline tool initiatives, such as STARD-AI[6] and TRIPOD-AI[7]. QUADAS-AI is being coordinated by a global project team and steering committee that consists of clinician scientists, computer scientists, epidemiologists, statisticians, journal editors, representatives of the EQUATOR Network[11], 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. ❐

Viknesh Sounderajah[1,2]✉, Hutan Ashrafian[1,2], Sherri Rose[3], Nigam H. Shah[4], Marzyeh Ghassemi[5], Robert Golub[6], Charles E. Kahn Jr[7], Andre Esteva[8], Alan Karthikesalingam[9], Bilal Mateen[10], Dale Webster[9], Dan Milea[11], Daniel Ting[11], Darren Treanor[12,13,14,15], Dominic Cushnan[16], Dominic King[1,17], Duncan McPherson[18], Ben Glocker[19], Felix Greaves[20], Leanne Harling[1,2], Johan

Table 1 | Examples of bias within AI DTA studies

Domain	Description	Biases
Patient selection	A description of included patients detailing previous testing, presentation, setting and the intended use of the index test.	In AI DTA studies, eligible patients are often excluded because of competing input data entry requirements (for example, image quality), which, themselves, are variably reported. As highlighted by the CONSORT-AI guidelines[8], there is a need to accurately characterize the source, size and quality of input data alongside clear patient eligibility criteria. Data source issues can negatively affect the performance and overall applicability of an index test. For example, to minimize research costs, there has been an increasing use of datasets from open-source repositories. Although this is a pragmatic option, many open-source datasets contain inadvertent duplication of data across repositories, erroneous labeling and incomplete patient demographic data. Manuscripts that report 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, in which 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 in which image formats and resolutions are homogenized for the purpose of training, is an essential step in AI workflows. However, either down- or up-scaling resolution may affect the ability of certain index tests to identify diagnostic features effectively. Moreover, the lack of image metadata can also preclude the ability to explore the dependence of an index test on specific data acquisition parameters—such as 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 on 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 several instances, as highlighted by Harris et al.[9], 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[10], studies that reported the performance of index tests to diagnose SARS-CoV-2 from chest X-rays did not routinely note the timing of the confirmatory PCR with reverse transcription test in relation to the imaging data. It is well understood that PCR with reverse transcription is a time-sensitive assay and failing to report this relationship considerably hinders the overall clinical validity of the study results.

Ordish[18], Jérémie F. Cohen[21], Jon Deeks[22], Mariska Leeflang[23], Matthew Diamond[24], Matthew D. F. McInnes[25], Melissa McCradden[26], Michael D. Abràmoff[27], Pasha Normahani[2], Sheraz R. Markar[2], Stephanie Chang[28], Xiaoxuan Liu[29,30,31], Susan Mallett[32], Shravya Shetty[9], Alastair Denniston[29,30,31], Gary S. Collins[33,34], David Moher[35], Penny Whiting[36], Patrick M. Bossuyt[23,37] ✉ and Ara Darzi[1,2,37] ✉

[1]Institute of Global Health Innovation, Imperial College London, London, UK. [2]Department of Surgery and Cancer, Imperial College London, London, UK. [3]Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA. [4]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [5]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Journal of the American Medical Association (JAMA), Chicago, IL, USA. [7]University of Pennsylvania, Philadelphia, Pennsylvania, PA, USA. [8]Salesforce Research, San Francisco, CA, USA. [9]Google Health, Palo Alto, CA, USA. [10]Wellcome Trust, London, UK. [11]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [12]Leeds Teaching Hospitals NHS Trust, Leeds, UK. [13]University of Leeds, Leeds, UK. [14]Department of Clinical Pathology, and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden. [15]Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden. [16]NHSX, London, UK. [17]Optum, London, UK. [18]Medicines and Healthcare Products Regulatory Agency, London, UK. [19]Faculty of Engineering, Department of Computing, Imperial College London, London, UK. [20]National Institute for Health and Care Excellence, London, UK. [21]Department of Pediatrics, Centre of Research in Epidemiology and Statistics, Inserm UMR 1153, Necker- Enfants Malades Hospital, Assistance Publique-Hôpitaux de Paris, Université de Paris, Paris, France. [22]Institute of Applied Health Research, University of Birmingham, Birmingham, UK. [23]Department of Epidemiology and Data Science, Amsterdam University Medical Centres, University of Amsterdam, Amsterdam, The Netherlands. [24]Food and Drug Administration, Silver Spring, MD, USA. [25]Departments of Radiology and Epidemiology, University of Ottawa, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada. [26]Department of Bioethics, The Hospital for Sick Kids, Toronto, Ontario, Canada. [27]Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, IA, USA. [28]Annals of Internal Medicine, American College of Physicians, Philadelphia, PA, USA. [29]Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [30]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [31]Health Data Research UK, London, UK. [32]Centre for Medical Imaging, University College London, London, UK. [33]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [34]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford,

UK. [35]*Ottawa Hospital Research Institute, Ottawa, Ontario, Canada.* [36]*Bristol Medical School, University of Bristol, Bristol, UK.* [37]*These authors jointly supervised this work: Patrick M. Bossuyt, Ara Darzi.* ✉*e-mail: viknesh.sounderajah08@imperial.ac.uk; p.m.bossuyt@amsterdamumc.nl; a.darzi@imperial.ac.uk*

Published online: 11 October 2021
https://doi.org/10.1038/s41591-021-01517-0

References

1. Benjamens, S., Dhunnoo, P. & Meskó, B. *npj Digit. Med.* **3**, 118 (2020).
2. Jayakumar, S. et al. *npj Digital Med.* (in the press).
3. Whiting, P. F. *Ann. Intern. Med.* **155**, 529 (2011).
4. Page, M. J. et al. *BMJ.* **372**, n71 (2021).
5. Yang, B. et al. *Open Science Framework* https://doi.org/10.17605/OSF.IO/HQ8MF (2018).
6. Sounderajah, V. et al. *Nat. Med.* **26**, 807–808 (2020).
7. Collins, G. & Moons, K. *Lancet* **393**, 1577–1579 (2019).
8. Liu, X. & Rivera, S. C. *Nat. Med.* **26**, 1364–1374 (2020).
9. Harris, M. et al. *PLoS One* **14**, e0226134 (2019).
10. Roberts, M. et al. *Nat. Mach. Intell.* **3**, 199–217 (2021).
11. The EQUATOR Network.Enhancing the QUAlity and Transparency Of Health Research; https://www.equator-network.org/ (accessed 27 September 2021).

Acknowledgements

Infrastructure support for this research was provided by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). G.S.C. is supported by the NIHR Biomedical Research Centre and Cancer Research UK (programme grant C49297/A27294). D.T. is funded by National Pathology Imaging Co-operative, NPIC (project no. 104687), supported by a £50 million investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, and managed and delivered by UK Research and Innovation (UKRI). F.G. is supported by the NIHR 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

V.S., S.R., N.H.S., M.G., R.G., C.E.K., X.L., G.S.C., D.W., A.E., H.A., D. Milea, D. McPherson, J.O., D. Treanor, J.F.C., M.L., M.M., M.D.F.M., M.D.A., S.M., P.W. and P.M.B. prepared the first draft of the manuscript. Critical edits, further direction and feedback have been attained from all co-authors (including A.K., B.M., D. Ting, D.C., D.K., F.G., L.H., J.D., M.D., P.N., S.M., S.C., S.S., A.D., D.M. and A.D.). The study described in the manuscript has been conceptualized, discussed and agreed upon between all co-authors.

Competing interests

A.K., S.S. and D.W. are employees at Google. A.D. and H.A. are employees at Flagship Pioneering UK Ltd. A.E. is an employee at Salesforce. DK is an employee at Optum. None of the other authors have any competing interests.