https://doi.org/10.1038/s41591-022-01987-w

Tackling bias in AI health datasets through the STANDING Together initiative

Check for updates

To the Editor — As of June 2022, a wide range of Artificial Intelligence (AI) as a Medical Device (AIaMDs) have received regulatory clearance internationally, with at least 343 devices cleared by the US Food and Drug Administration (FDA)[1]. Despite the enormous potential of AIaMDs, their rapid growth in healthcare has been accompanied by concerns that AI models may learn biases engrained in medical practice and exacerbate health inequalities. This has been exemplified by several AI systems that have shown the ability of algorithms to systematically misrepresent and exacerbate health problems in minority groups[2,3]. This raises concerns that, without appropriate safeguarding, AI models may perpetuate existing health inequality and mistrust.

Tackling bias in AI requires a multifaceted approach. A recent report by the US National Institute of Standards and Technology on bias in AI emphasized that algorithmic development does not occur by engineering decisions alone, but embeds a myriad of values and behaviors within the data and the humans who interact with them. The report calls for a sociotechnical approach that considers how different biases interact and the social contexts within which AI systems are built and used[4]. Although there is an expanding field of research dedicated to fairness in machine learning, many AIaMDs that receive regulatory clearance have not appropriately accounted for biases that disadvantage certain populations. There are also ethical challenges around algorithmic fairness methods (computational techniques that seek to ensure outputs are not unjustifiably influenced by bias), given that these methods are aimed at making predictions fair, rather than enabling the fair treatment of individuals[5]. Furthermore, current approaches to satisfy regulatory requirements are focused on aggregate-level performance, which can mask stratification across subpopulations.

One major source of bias is the data that underpin AI systems. It is often necessary to train models with large quantities of data, which means datasets are often sourced to prioritize sample size. There are concerns that many health datasets do not adequately represent minority groups; however, the extent of this problem is unknown because many datasets do not provide demographic information, such as on ethnicity and race. Publicly available datasets for skin cancer and eye imaging have shown inconsistent and incomplete demographic reporting, and are disproportionately collected from a small number of high-income countries[6,7]. For skin cancer datasets, the reporting of key demographic information, such as ethnicity and skin tone, even when clinically relevant, was only present in 2% of datasets[7].

Under-representation in datasets can affect the fairness of AI systems by two principal means. During AI development, under-representation within training datasets can negatively affect model performance for under-represented groups[3]. A lack of diversity within the training data risks poor generalizability of model performance after deployment. During evaluation, under-representation within test datasets increases the uncertainty of performance in that group due to small sample sizes, which reduces the likelihood of detecting under-performance. Therefore, under-representation not only creates models that under-perform within minority populations, but also hampers the ability to detect this bias. Furthermore, under-representation in datasets may result in exclusion of populations from the intended use altogether, thereby creating AI systems licensed for only certain groups within society. Even when datasets are inclusive, additional issues can compound bias. Structural inequities can manifest in datasets through the actions of clinical and data curation teams, who are responsible for recording, selecting, labelling and aggregating data, based on assumptions that reflect hegemonic social attitudes. Addressing the consequences of structural biases requires a wider consideration of the dataset: how and why it was created; the setting in which the data was collected and by whom; the extent to which the data reflect broader structural biases and axes of injustice; the inclusion and exclusion criteria; and how measurements, observations and labels were constructed. These concerns have motivated calls for better documentation practices and the creation of tools such as 'Datasheets for Datasets' and 'Healthsheets'[8,9].

The aforementioned problems are becoming increasingly recognized by regulators of medical devices. In October 2021, The US FDA, Health Canada and the UK Medicines and Healthcare products Regulatory Agency (MHRA) jointly published 10 guiding principles for good machine learning practice. This specifically states that data should be representative of the intended population in order to "manage any bias, promote appropriate and generalizable performance across the intended patient population, assess usability and identify circumstances where the model may underperform"[10]. Commitment to identify and mitigate bias by medical regulators is an important step in the right direction; however, there is a lack of evidence that these principles are adopted by AIaMD manufacturers. Without specific consensus on how to assess the appropriateness of datasets, it is unclear what constitutes best practice regarding the use of health data in AI to promote fairness and equity.

To tackle this problem, we are announcing the development of the STANDING Together (standards for data diversity, inclusivity and generalisability) initiative. This is an international, consensus-based initiative that aims to develop recommendations for the composition (who is represented) and reporting (how they are represented) of datasets that underpin medical AI systems. We will engage patients and the public, clinicians and academic researchers across biomedical, computational and social sciences, industry experts, regulators and policy-makers. The standards will represent the culmination of a multiphase evidence generation process, which consists of: dataset mapping reviews to assess limitations in health datasets across different diseases with regard to diversity and inclusivity; interviews with dataset curators to explore the barriers and challenges to ensuring diversity and inclusivity within health datasets; a modified Delphi consensus study to finalize the content that will feature in these recommendations; and an extensive multi-stakeholder piloting phase.

The resulting standards will support informed decision-making for those who strive to engineer and implement fair and safe AI systems in healthcare. STANDING Together will be the first in a line of work through which stakeholders can determine what demographic data is collected and how it is represented in datasets. The findings will motivate curators of health datasets to prioritize diversity and inclusiveness as we all seek to build and invest in health datasets of the future. We hope that this initiative will enable the availability of more inclusive data to promote responsible AI in healthcare, and in the long term, better health outcomes for all.

The modified Delphi consensus study will begin in September 2022 and the final standards will be published in 2023. We welcome those with expertise in AI, health data science and health inequalities to participate (through https://www.datadiversity.org/delphi or by contacting contact@datadiversity.org).

Shaswath Ganapathi[1], **Jo Palmer**[2], **Joseph E. Alderman**[2,3], **Melanie Calvert**[4,5,6,7,8], **Cyrus Espinoza**[9], **Jacqui Gath**[9,10], **Marzyeh Ghassemi**[11], **Katherine Heller**[12], **Francis Mckay**[13], **Alan Karthikesalingam**[14], **Stephanie Kuku**[15,16], **Maxine Mackintosh**[17], **Sinduja Manohar**[18], **Bilal A. Mateen**[19,20], **Rubeta Matin**[21], **Melissa McCradden**[22,23], **Lauren Oakden-Rayner**[24], **Johan Ordish**[25], **Russell Pearson**[25], **Stephen R. Pfohl**[12], **Negar Rostamzadeh**[26], **Elizabeth Sapey**[2,3], **Neil Sebire**[18,27], **Viknesh Sounderajah**[28,29], **Charlotte Summers**[30,31], **Darren Treanor**[32,33,34,35], **Alastair K. Denniston**[2,3,4,6,18] and **Xiaoxuan Liu**[2,3,4] ✉

[1] College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [2] University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [3] Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [4] Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK. [5] Centre for Patient Reported Outcome Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK. [6] NIHR Birmingham Biomedical Research Centre, University of Birmingham, Birmingham, UK. [7] NIHR Surgical Reconstruction and Microbiology Research Centre, University of Birmingham, Birmingham, UK. [8] NIHR Applied Research Collaborative West Midlands University of Birmingham, Birmingham, UK. [9] Patient Partner, Birmingham, UK. [10] Patient Partner, Sheffield, UK. [11] Department of Electrical Engineering and Computer Science; Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [12] Google Research, Mountain View, California, USA. [13] The Ethox Centre and the Wellcome Centre for Ethics and Humanities, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [14] Google Research, London, UK. [15] Institute of Women's Health, University College London, London, UK. [16] Hardian Health, London, UK. [17] Genomics England, London, UK. [18] Health Data Research, London, UK. [19] Institute of Health Informatics, University College London, London, UK. [20] The Wellcome Trust, London, UK. [21] Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [22] Department of Bioethics, Hospital for Sick Children, Toronto, Ontario, Canada. [23] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [24] Australian Institute for Machine Learning, University of Adelaide, Adelaide, South Australia, Australia. [25] Medicines and Healthcare Products Regulatory Agency, London, UK. [26] Google Research, Montreal, Canada. [27] Great Ormond Street Hospital for Children, London, UK. [28] Institute of Global Health Innovation, Imperial College London, London, UK. [29] Department of Surgery and Cancer, Imperial College London, London, UK. [30] Wolfson Lung Injury Unit, Heart and Lung Research Institute, University of Cambridge, Cambrdige, UK. [31] Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [32] Leeds Teaching Hospitals NHS Trust, Leeds, UK. [33] University of Leeds, Leeds, UK. [34] Department of Clinical Pathology, and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden. [35] Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden.
✉e-mail: x.liu.8@bham.ac.uk

Published online: 26 September 2022

References

1. FDA Center for Devices & Radiological Health. https://go.nature.com/3AG0McN (2021).
2. Obermeyer, Z. et al. *Science* **366**, 447–453 (2019).
3. Seyyed-Kalantari, L. et al. *Nat. Med.* **27**, 2176–2182 (2021).
4. Schwartz, R. et al. National Institute of Standards and Technology; https://go.nature.com/3Q6rjpj (2022).
5. McCradden, M. D. et al. *Lancet Digit Health* **2**, e221–e223 (2020).
6. Khan, S. M. et al. *Lancet Digit Health* **3**, e51–e66 (2021).
7. Wen, D. et al. *Lancet Digit Health* **4**, e64–e74 (2022).
8. Rostamzadeh, N. et al. *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency* https://doi.org/10.1145/3531146.3533239 (2022).
9. Gebru, T. et al. Preprint at https://doi.org/10.48550/arXiv.1803.09010 (2018).
10. Medicines and Healthcare products Regulatory Agency. https://go.nature.com/3RsijvS (2021).

Acknowledgements

This project is funded by The NHS AI Lab at the NHS Transformation Directorate and the Health Foundation and managed by the National Institute for Health and Care Research (AI_HI200014). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS Transformation Directorate, the Health Foundation or the National Institute for Health and Care Research.

Competing interests

K.H., A.K., N.R. and S.R.P. are employees of Google. S.K. is a consultant for Hardian Health. D.T. and F.M. are funded by National Pathology Imaging Co-operative (NPIC, Pproject no. 104687), which is supported by a £50 million 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). X.L., A.K.D., J.E.A. and J.P. are funded by NIHR, the NHS Transformation Directorate and the Health Foundation (AI_HI200014). M.J.C. is Director of the Birmingham Health Partners Centre for Regulatory Science and Innovation, Director of the Centre for the Centre for Patient Reported Outcomes Research and is a National Institute for Health and Care Research (NIHR) Senior Investigator. M.J.C. receives funding from the NIHR, UK Research and Innovation (UKRI), NIHR Birmingham Biomedical Research Centre, the NIHR Surgical Reconstruction and Microbiology Research Centre, NIHR ARC West Midlands, UK SPINE, European Regional Development Fund – Demand Hub and Health Data Research UK at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Innovate UK (part of UKRI), Macmillan Cancer Support, UCB Pharma, Janssen, GSK and Gilead, has received personal fees from Astellas, Aparito Ltd, CIS Oncology, Takeda, Merck, Daiichi Sankyo, Glaukos, GSK and the Patient-Centered Outcomes Research Institute (PCORI) outside the submitted work; a family member of M.J.C. owns shares in GSK. ES receives research funding from UKRI (MR/V033654/1 and MR/S002782/1), the British Lung Foundation, and Alpha 1 Foundation and NIHR. C.S. receives research funding from the National Institute for Health and Care Research (NIHR133788), UKRI (MR/P502091/1 and MR/X005070/1), the Wellcome Trust, and the NIHR Cambridge Biomedical Research Centre (BRC1215-20014). All other authors declare no conflicts.