frontiers | Frontiers in Digital Health

TYPE Opinion
PUBLISHED 25 February 2025
DOI 10.3389/fdgth.2025.1540982

OPEN ACCESS

EDITED BY
Stephen Fashoto,
University of Eswatini, Eswatini

REVIEWED BY
Marcin Golec,
Heidelberg University Hospital, Germany

*CORRESPONDENCE
Sahar Abdulrahman
✉ Sahar.x.abdulrahman@gsk.com

RECEIVED 06 December 2024
ACCEPTED 04 February 2025
PUBLISHED 25 February 2025

CITATION
Abdulrahman S and Trengove M (2025) Levelling up as a fair solution in AI enabled cancer screening.
Front. Digit. Health 7:1540982.
doi: 10.3389/fdgth.2025.1540982

Levelling up as a fair solution in AI enabled cancer screening

Sahar Abdulrahman* and Markus Trengove

GSK.ai, GSK, King's Cross, London, United Kingdom

KEYWORDS

AI, equity, fairness, screening, cancer, inequalities

1 Introduction

With the proliferation of artificial intelligence (AI) enabled tools in healthcare, clinicians have raised concerns about the potential for bias and subsequent negative impacts on underrepresented groups (1). The causes of bias in model deployment are multifaceted and can occur throughout the model development process (2). A well-recognized example within model training includes unrepresentative datasets that limit model generalizability in real-world populations (3). Whilst AI may outperform current standards of care for well-represented groups, models can perform worse for under-represented groups (4–9). Diversifying datasets is the obvious solution to bias caused by homogenous training data, however data collection is a long term project that may take years or even decades to acquire (10). The dilemma for policymakers currently is that releasing unfair tools can harm under-represented groups, whilst withholding them would cause significant welfare opportunity costs for well-represented groups. To address this problem, bioethicists like Vandersluis and Savalescu (11) have suggested alternate deployment strategies such as "selective deployment", which would deploy AI tools only to well-represented groups. This incurs an obvious fairness cost. The issue of fairness in diagnostic testing is not specific to AI applications and debates also remain ongoing in the field of mainstream medicine on how to address this challenge (12, 13). By examining the case study of faecal immunochemical test (FIT) screening, which has been shown to perform more effectively for male patients in detecting bowel cancer, this paper supports the use of sex adjustment to "level up" female patients (14). Through this example within mainstream medicine, lessons learned from real-world policy can be transferred to clinical AI deployment which will be illustrated using a parallel case study of AI-assisted breast cancer screening.

2 Bowel cancer screening

National Health Service (NHS) England introduced FIT screening for bowel cancer detection in 2019, see Figure 1 for summary of clinical workflow (15). FIT tests are stool samples that measure the concentration of blood in the stool, and if above a specified threshold triggers referral for further investigation which is usually a colonoscopy. In the UK, the National Screening Committee (NSC) set the FIT threshold (120 μg/g) based on cost-effectiveness analysis, where effectiveness is determined by Quality-Adjusted Life Years (QALY) gained, and system capacity (16). A lower FIT threshold results in more "positive" tests, with a subsequent greater rate of unnecessary colonoscopies. Conversely, higher FIT thresholds will result in a lower burden on colonoscopy constraints at the risk of more missed cancers. There is increasing evidence that FITs perform worse for female patients, who have a lower median faecal blood measurement than males (17). For each FIT threshold, cancer detection rates have been found to be lower in the female subgroup (18). Despite this, the UK continues to use universal FIT thresholds in contrast to some other

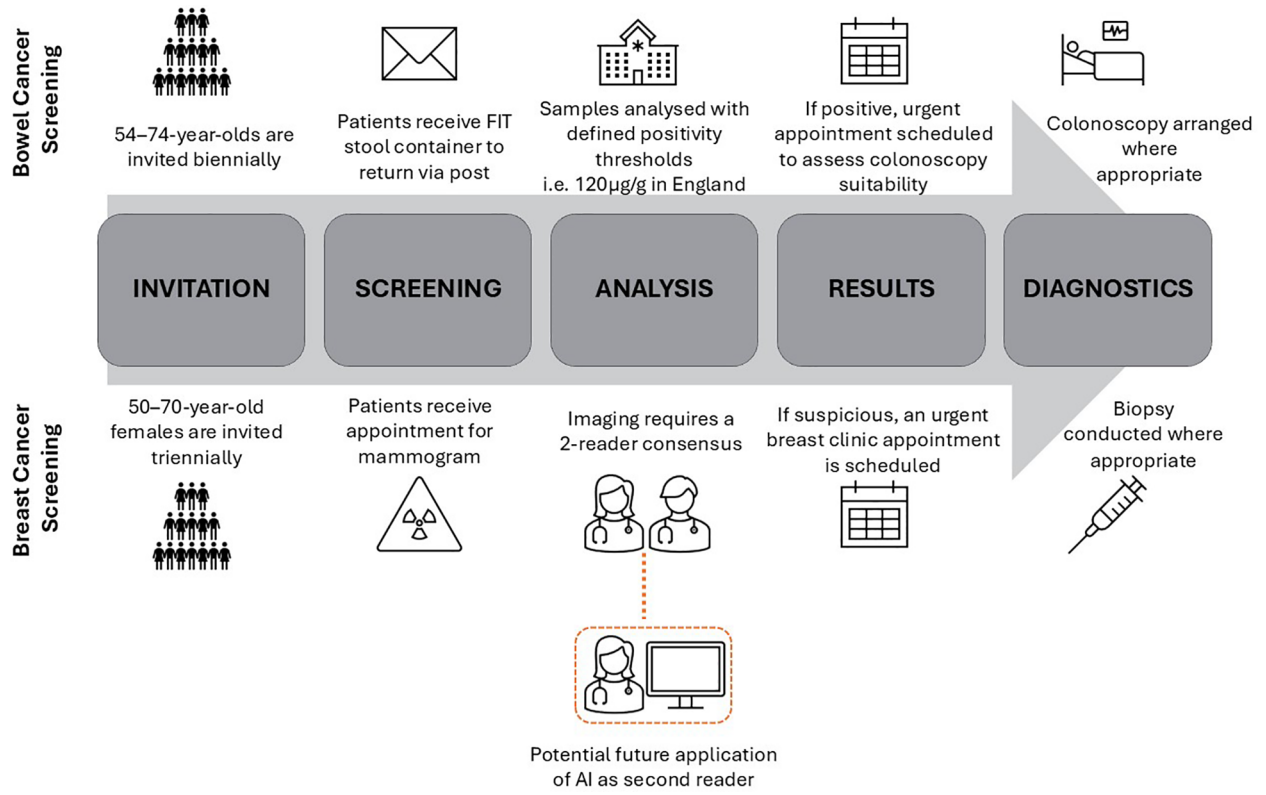

FIGURE 1
NHS clinical workflows of parallel case studies (bowel and breast national screening).

countries who have adopted sex-adjusted thresholds resulting in a positive test at a lower blood concentration in female patients (19). For example, Sweden's thresholds for positivity of 40 µg/g and 80 µg/g for females and males respectively, has resulted in more equal proportions of cancer detection in subgroups but at the cost of higher rates of negative colonoscopies (20). Similar trends have also been seen in Finland who have sex-adjusted thresholds (21).

3 Discussion

A cost-benefit analysis of sex-adjusted FIT thresholds will be conducted. The aim of which is to assess the effectiveness of "levelling up" through sex adjustment in ensuring equitable health outcomes when "unfair" diagnostic tests are being utilised. Following this, using a parallel case study of breast cancer screening, the way in which levelling up can be applied to AI deployments, such as AI assisted mammogram interpretation, will be explored. These case studies will illustrate the usefulness of transfer learning from mainstream medicine to the emerging field of algorithmic fairness.

3.1 Cost-benefit analysis

In the case of bowel screening, a lower FIT threshold for female patients is the fairest strategy for maximal utility. From a public health perspective, increased detection of cancer for female patients at levels similar to men has the potential to reduce overall bowel cancer mortality and morbidity (22). Economically, by reducing the false negative rate in female patients, earlier detection of cancer can be more cost effective as earlier presentations will be more amenable to treatment which is particularly important in a publicly funded health system (23). In turn, overall health service costs are reduced by first line treatments, and reduced social care costs associated with advanced cancer. Health gains will vary between countries due to differences in underlying population risk, but evidence from Sweden shows that nearly 25% of female patients who would have been classified as negative by universal thresholds, were subsequently diagnosed with bowel cancer as a result of the lower sex-adjusted thresholds (19). From an ethical standpoint, lower thresholds for female patients acknowledges that universal processes may be suboptimal in ensuring fair outcomes, particularly in medicine where much of the evidence base is grounded on white male normativity (24).

The costs of levelling up disadvantaged subgroups centre on the impacts of more false positives. Firstly, the clinical risk posed by higher rates of false positives will differ depending on application, and whilst colonoscopies are not completely free of harm they do constitute a relatively lower risk intervention. For example, a randomised trial exploring the effect of colonoscopy screening found that of approximately 12,000 patients who had a colonoscopy, there were no bowel perforations or deaths within the 30 days post procedure (25). Furthermore, specialist

nurses screen all patients with a positive FIT test before colonoscopy to ensure they are fit enough for the procedure as a further safety net to mitigate harm (26). Secondly, a significant cost in levelling up is the effect on system capacity as healthcare providers may not be able to provide necessary colonoscopies due to constraints of unit space, equipment and qualified personnel (27). Some may argue that in this case it would be equitable to increase the threshold for male patients whilst keeping the female threshold the same as this could lead to equivalent performance between subgroups without exceeding existing colonoscopy capacity. Although it may lead to more similar outcomes, it is widely accepted that down-levelling the male group by increasing the existing missed cancer rate would be unethical. Instead, healthcare policymakers must mandate subgroup analysis prior to deployment in a responsible by design approach, so that appropriate thresholds can be set with both subgroup performance and system constraints being considered.

3.2 Healthcare specific challenges

Levelling up in healthcare presents unique challenges specific to clinical medicine. Although this paper advocates for adjustments to mitigate for poor performance in subgroups, it acknowledges that there is a need for post-deployment evaluation due to the complex nature of disease manifestation. Data drift refers to changes in the properties of data over time from what was used in model training (28). Although FIT tests are not enabled by AI, a phenomena similar to data drift can occur whereby changes in disease can mean FIT threshold are no longer appropriate; for example, bowel cancer presenting more in younger patients and declining incidence in men (29). Therefore, thresholds should not be eternally fixed and regular post-deployment evaluation should be conducted to ensure that thresholds are continuing to be fit for purpose and meeting the ever-changing needs of patients.

Whilst FIT testing disadvantages female patients, there are other subgroups who are also not well-represented in both clinical trial and training data, such as non-white racial groups, who would benefit from levelling up in other contexts. Race-adjustment may be more difficult to adopt than sex-adjustment due to controversies surrounding race-based medicine stemming from historically exploitative practices such as Sims' experimentation on enslaved black women (30). This paper acknowledges that race is social construct and is also critical of the historic underpinnings of how race categories were and continue to be defined (31). However, race-adjustment can be a useful tool in addressing health inequality in instances where it is used to uplevel groups who receive inadequate care due to system failures, in part due to the impacts of systemic racism, rather than to propagate the belief of innate biological differences. When levelling up poorly performing subgroups with adjustment, transparency will be critical in ensuring understanding and trust in healthcare providers, particularly in groups who have faced historic injustice.

3.3 Conditions for levelling up

Though adjustment can be a useful tool to mitigate for differential performance in subgroups, this should not be seen as one-size-fits-all solution. Rather, adjustment is intended to add to existing research on strategies for deploying such models, allowing for a comprehensive guide that policymakers can draw from. There are specific conditions under which adjustment is the most suitable approach. For this mitigation strategy to be effective there needs to be subgroup analysis that identifies a negative bias, specifically an underdiagnosis. Furthermore, adjustment will result in higher rates of false positive results for subgroups and deploying teams must understand the clinical sequalae of over referral. A false positive in different clinical contexts will have different repercussions, which can also be the case with the same application deployed across separate NHS trusts who have varying guidelines. Adjustment is preferred in contexts where there is a low-risk intervention, with high gain such as in the case of cancer screening. Next, workflows where there is a human-in-the-loop will mitigate harm of over-referral, such as the specialist nurse contact to assess fitness for colonoscopy in FIT testing.

An example of how these conditions can apply to clinical AI deployment includes the use of AI in the parallel case study of breast cancer screening. Similarly to FIT testing, mammograms are offered as a screening test in a national cancer screening programme in the NHS, see Figure 1 for clinical workflow. The use of AI in imaging, also known as computer vision, is the most popular application of AI in the health service (32). Despite the NSC finding a lack of evidence to introduce AI in NHS breast screening, countries like Sweden have already begun trialing AI as a second reader of mammograms in prospective studies (33, 34). Recent research has highlighted that a commercially available model diagnosing suspicious lesions from mammogram images overpredicts suspicious lesions in the images of black patients (35). Despite the concerns that this research has raised, there is a context in which this could be a harm mitigation strategy. An intentional higher false positive rate in black patients could be an example of levelling up if there was an initial underdiagnosis bias in this subgroup. Levelling up would be suitable given the high gain of possible cancer detection and low risk due to the double read requirement on mammograms in national screening (36). The existing workflow acts to reduce risk by ensuring one of the readers is a clinician-in-the-loop who can query AI diagnoses and seek a third reader opinion if necessary. Furthermore, if this safety net fails (i.e., both AI and human reader wrongly classify as suspicious), an urgent specialty review is organised to decide if a biopsy is necessary further lowering the risk to patients.

Levelling up doesn't solve the reasons why models may perform differentially, but does offer a solution in how to mitigate for harm through use of the clinical workflow. The causes of algorithmic bias are multifaceted and can happen at each point in the model development pipeline. As such, cross functional teams including developers, clinicians and researchers must attempt to elicit such causes and act together

to highlight possible interventions to counteract preventable root causes. An example of such interventions includes initiatives to engage with underrepresented groups in data collection efforts and emerging techniques such as the use of synthetic data to diversify datasets (37–39).

4 Summary

In summary, levelling up can be an approach that safely balances fairness and utility when certain conditions are met. The parallel case studies highlights the usefulness of transfer learning from mainstream medicine, where solutions to unfair diagnostics have a real-world evidence base, to clinical AI. Whilst levelling up is a useful strategy to mitigate harm, it is essential that there remains a focus on addressing preventable root causes of algorithmic bias.

Author contributions

SA: Writing – original draft, Writing – review & editing. MT: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was completed as part of the GSK.ai fellowship within the Responsible AI department. The views expressed in the paper are the author's own, and do not necessarily reflect of views of GSK or GSK.ai.

Conflict of interest

SA and MT are current employees and shareholders at GSK, which is a pharmaceutical company that conducts AI research. SA and MT have a working relationship with Robert Vandersluis at GSK who is referenced in the manuscript.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Whitehead M, Carrol E, Kee F, Holmes C. Making the invisible visible: what can we do about biased AI in medical devices? *Br Med J*. (2023):1893. doi: 10.1136/bmj.p1893

2. Tejani AS, Retson TA, Moy L, Cook TS. Detecting common sources of AI bias: questions to ask when procuring an AI solution. *Radiology*. (2023) 307(3):e230580. doi: 10.1148/radiol.230580

3. Kamikubo R, Wang L, Marte C, Mahmood A, Kacorri H. Data representativeness in accessibility datasets: a meta-analysis. In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. Athens Greece: ACM*; (2022). p. 1–15. doi: 10.1145/3517428.3544826 (accessed 9 January 2025).

4. Goankar B, Cook K, Macyszyn L. Ethical issues arising due to bias in training A.I. algorithms in healthcare and data sharing as a potential solution. *AI Ethics Journal*. (2020) 1(2):1–9. doi: 10.47289/AIEJ20200916

5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366(6464):447–53. doi: 10.1126/science.aax2342

6. Salinas MP, Sepúlveda J, Hidalgo L, Peirano D, Morel M, Uribe P, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *NPJ Digit Med*. (2024) 7(1):125. doi: 10.1038/s41746-024-01103-x

7. Stanley EAM, Wilms M, Mouches P, Forkert ND. Fairness-related performance and explainability effects in deep learning models for brain image analysis. *J Med Imaging*. (2022) 9(06):061102-1–17. doi: 10.1117/1.JMI.9.6.061102

8. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. (2021) 27(12):2176–82. doi: 10.1038/s41591-021-01595-0

9. Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Front Cardiovasc Med*. (2022) 9:859310. doi: 10.3389/fcvm.2022.859310

10. Andrus M, Spitzer E, Brown J, Xiang A. 'What We Can't Measure, We Can't Understand': Challenges to Demographic Data Procurement in the Pursuit of Fairness. (2020). doi: 10.48550/ARXIV.2011.02282

11. Vandersluis R, Savulescu J. The selective deployment of AI in healthcare: an ethical algorithm for algorithms. *Bioethics*. (2024) 38(5):391–400. doi: 10.1111/bioe.13281

12. Kadambi A. Achieving fairness in medical devices. *Science*. (2021) 372(6537):30–1. doi: 10.1126/science.abe9195

13. Figueroa ML, Hiemstra LA. How do we treat our male and female patients? – A primer on gender-based health care inequities. *J ISAKOS*. (2024) 9(4):774–80. doi: 10.1016/j.jisako.2024.04.006

14. Ribbing Wilén H, Saraste D, Blom J. Interval cancers in a population-based screening program for colorectal cancer with gender-specific cut-off levels for fecal immunochemical test. *J Med Screen*. (2022) 29(3):156–65. doi: 10.1177/09691413221085218

15. NICE. What is the NHS bowel screening programme in the UK? Available online at: https://cks.nice.org.uk/topics/bowel-screening/background-information/the-nhs-bowel-screening-programme/ (accessed April 10, 2024).

16. Whyte S, Thomas C, Kearns B, Webster M, Chilcott J. Optimising Bowel Cancer Screening Phase 1: Optimising the cost effectiveness of repeated FIT screening and screening strategies combining bowel scope and FIT screening (2018). Available online at: https://view-health-screening-recommendations.service.gov.uk/document/318/download (accessed October 10, 2024).

17. White A, Ironmonger L, Steele RJC, Ormiston-Smith N, Crawford C, Seims A. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC Cancer*. (2018) 18(1):906. doi: 10.1186/s12885-018-4786-7

18. Van Turenhout ST, Oort FA, Van Der Hulst RW, Visscher AP, Terhaar Sive Droste JS, Scholten P, et al. Prospective cross-sectional study on faecal immunochemical tests: sex specific cut-off values to obtain equal sensitivity for colorectal cancer? *BMC Gastroenterol*. (2014) 14(1):217. doi: 10.1186/s12876-014-0217-7

19. Ribbing Wilén H, Saraste D, Blom J. Gender-specific cut-off levels in colorectal cancer screening with fecal immunochemical test: a population-based study of colonoscopy findings and costs. *J Med Screen*. (2021) 28(4):439–47. doi: 10.1177/09691413211020035

20. Ribbing Wilén H, Blom J. Interval cancer after two rounds of a Swedish population-based screening program using gender-specific cut-off levels in fecal immunochemical test. *J Med Screen*. (2024) 31(1):8–14. doi: 10.1177/09691413231185722

21. Sarkeala T, Färkkilä M, Anttila A, Hyöty M, Kairaluoma M, Rautio T, et al. Piloting gender-oriented colorectal cancer screening with a faecal immunochemical test: population-based registry study from Finland. *BMJ Open*. (2021) 11(2): e046667. doi: 10.1136/bmjopen-2020-046667

22. Chen C, Stock C, Hoffmeister M, Brenner H. Public health impact of colonoscopy use on colorectal cancer mortality in Germany and the United States. *Gastrointest Endosc*. (2018) 87(1):213–221.e2. doi: 10.1016/j.gie.2017.04.005

23. Lew JB, St John DJB, Xu XM, Greuter MJE, Caruana M, Cenin DR, et al. Long-term evaluation of benefits, harms, and cost-effectiveness of the national bowel cancer screening program in Australia: a modelling study. *Lancet Public Health*. (2017) 2(7): e331–40. doi: 10.1016/S2468-2667(17)30105-6

24. Plaisime MV, Jipguep-Akhtar MC, Belcher HME. 'White people are the default': a qualitative analysis of medical trainees' perceptions of cultural competency, medical culture, and racial bias. *SSM – Qual Res Health*. (2023) 4:100312. doi: 10.1016/j.ssmqr.2023.100312

25. Bretthauer M, Løberg M, Wieszczy P, Kalager M, Emilsson L, Garborg K, et al. Effect of colonoscopy screening on risks of colorectal cancer and related death. *N Engl J Med*. (2022) 387(17):1547–56. doi: 10.1056/NEJMoa2208375

26. NHS. Bowel cancer screening. https://www.nhs.uk/conditions/bowel-cancer-screening/#:~:text=You%20will%20get%20a%20letter,helpline%20on%200800%20707%206060 (accessed November 27, 2014).

27. McFerran E, O'Mahony JF, Naber S, Sharp L, Zauber AG, Lansdorp-Vogelaar I, et al. Colorectal cancer screening within colonoscopy capacity constraints: can FIT-based programs save more lives by trading off more sensitive test cutoffs against Longer screening intervals? *MDM Policy Pract*. (2022) 7(1):23814683221097064. doi: 10.1177/23814683221097064

28. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol*. (2023) 96(1150):20220878. doi: 10.1259/bjr.20220878

29. Cancer Research UK. Bowel cancer incidence statistics. Available online at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/risk-factors (accessed October 18, 2024).

30. Dudley R. Honoring the enslaved African American foremothers of modern women's health: meditations on 40 years of black feminist praxis. *Med Anthropol Q*. (2023) 38(4):445–61. doi: 10.1111/maq.12836

31. Braveman P, Parker Dominguez T. Abandon "race." focus on racism. *Front Public Health*. (2021) 9:689462. doi: 10.3389/fpubh.2021.689462

32. Zhu S, Gilbert M, Chetty I, Siddiqui F. The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: an analysis of the characteristics and intended use. *Int J Med Inf*. (2022) 165:104828. doi: 10.1016/j.ijmedinf.2022.104828

33. Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *Lancet Digit Health*. (2023) 5(10):e703–11. doi: 10.1016/S2589-7500(23)00153-X

34. Taylor-Phillips S, Seedat F, Kijauskaite G, Marshall J, Halligan S, Hyde C, et al. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *Lancet Digit Health*. (2022) 4(7):e558–65. doi: 10.1016/S2589-7500(22)00088-7

35. Nguyen DL, Ren Y, Jones TM, Thomas SM, Lo JY, Grimm LJ, et al. Patient characteristics impact performance of AI algorithm in interpreting negative screening digital breast tomosynthesis studies. *Radiology*. (2024) 311(2):e232286. doi: 10.1148/radiol.232286

36. NHS England. Breast screening: guidance for image reading (2024). Available online at: https://www.gov.uk/government/publications/breast-screening-guidance-for-image-reading/breast-screening-guidance-for-image-reading#:~:text=1%20Double%20reading%20of%20mammograms,be%20paired%20with%20experienced%20readers (accessed October 18, 2024).

37. Ganapathi S, Palmer J, Alderman JE, Calvert M, Espinoza C, Gath J, et al. Tackling bias in AI health datasets through the STANDING together initiative. *Nat Med*. (2022) 28(11):2232–3. doi: 10.1038/s41591-022-01987-w

38. Arora A. Synthetic data: the future of open-access health-care datasets? *Lancet*. (2023) 401(10381):997. doi: 10.1016/S0140-6736(23)00324-0

39. Arora A, Wagner SK, Carpenter R, Jena R, Keane PA. The urgent need to accelerate synthetic data privacy frameworks for medical research. *Lancet Digit Health*. (2024) 7(2):e157–60. doi: 10.1016/S2589-7500(24)00196-1