Commentary

Journal of the Royal Society of Medicine; 2023, Vol. 116(8) 260–262
DOI: 10.1177/01410768231187734

Addressing ethnic and global health inequalities in the era of artificial intelligence healthcare models: a call for responsible implementation

Mohammad R Ali[1], **Claire A Lawson**[1], **Angela M Wood**[2] **and Kamlesh Khunti**[3]

[1]Department of Cardiovascular Sciences, University of Leicester, Leicester, LE1 7RH, UK
[2]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, CB2 0SR, UK
[3]Ethnic Health Research, Diabetes Research Centre, University of Leicester, Leicester, LE5 4PW, UK
Corresponding author: Mohammad R Ali. Email: mra30@leicester.ac.uk

Artificial intelligence (AI) tools are being used more frequently in healthcare, particularly in medical applications such as radiology and risk prediction.[1] With this comes the palpable excitement to democratise healthcare using new AI models such as ChatGPT. However, we must tread carefully and act urgently to ensure that these large language models (LLMs) and their iterations do not exacerbate recently highlighted ethnic and global health inequalities.

The biases in AI systems often stem from their training data. For LLMs, this training process starts with 'pre-training': a large corpus of data scraped and ingested by the model from disparate sources such as websites and scientific literature. Using these data, a technique called 'self-supervised learning' predicts words based on the preceding word it has written. These models can then be fine-tuned for specific tasks/domains.[2] However, due to potential risks it poses to societal safety, concerns have been raised about opacity of the data sources used to train LLMs.[3] A particularly dangerous prospect for healthcare LLMs is their susceptibility to 'hallucinations' – creating ungrounded, subtly incorrect information without self-awareness.[4,5]

Given the nature of these training processes, there is a risk that health inequalities may become even more entrenched, particularly for ethnic minorities, due to systemic structural and societal biases in healthcare research perpetuated when training LLMs. Evidence depicts that ethnicity data are often missing[6] or unrecorded due to privacy concerns.[7] Compounding this is the inability to define an accurate scale of how little ethnic minorities participate in research.[8] The broader generalisability of research findings is already concerning as participants are less representative than those receiving the evaluated interventions in trials.[9] The scarcity of published research and therefore smaller sample size with ethnic minorities will mean 'less certainty' statistically and, thus, less accuracy regarding effect sizes of interventions. Furthermore, this disproportionately lower representation of ethnic minorities in research has evidence of causing harm, e.g. by creating ineffective drug treatments[10] or treatment guidance which could be regarded as racist.[11] It is widely accepted that a differential risk is associated with being from an ethnic minority background across many disease groups.[12] If the published literature already contains biases and less precision, it is logical that future AI models will maintain and further exacerbate them.

Beyond the issues of ethnic inequalities, there is a significant concern that health inequalities could worsen in low- and middle-income countries (LMICs), which suffer from most chronic disease without the requisite resources. AI models are primarily developed in wealthier nations like the USA and Europe, collectively known as the Global North, and a significant disparity in research and development exists between high- and low-income countries.[13] Most published research does not prioritise the needs of those in the LMICs with their unique health challenges, particularly around healthcare provision. LLMs, using published literature, may provide advice based on the corpus of data trained on populations wholly different from those in LMICs. Inevitably, the outputs of these and any future AI models being widely generalisable and truly inclusive will be limited.

While crucial to acknowledge these potential difficulties, it is equally important to focus on solutions. We must exercise caution acknowledging we cannot and should not stem the flow of progress.

Nevertheless, we ask researchers to consider the broader generalisability of AI models including LLMs. Several actions need to be implemented to overcome potentially exacerbating health inequalities.

First, AI models should clearly describe their body of medical data used in model development's 'pre-training' and refinement stages, including ethical approvals. Second, work is needed to address ethnic health inequalities in research. This includes recommendations to improve recruitment and recording of ethnicity information.[8,9] Encouraging wider participation and recording of ethnicity data will enable AI models to be trained on representative populations. Third, we must ensure that the data used to train the AI model are adequately representative while considering potential model biases and its accuracy. The model must consider key factors, including ethnicity, age, sex and socioeconomic factors. Fourth, using appropriate methodologies, further research is required to understand the generalisability of LLMs and other AI models in ethnically diverse populations. Researchers should consider recalibrating models before using them in ethnically diverse populations. Moreover, when utilising published research, it is advisable to assess the risk of bias[14] and research quality using tools like GRADE.[15] Given their reliance solely on the training data coupled with their inability to reason, these models have limited capability in discerning between 'good' and 'poor' research. By ensuring that only high-quality research is used for training AI models, we can reduce the risk of providing incorrect answers.

In summary, LLMs and future AI models have the potential for transforming healthcare. However, caution is warranted before these models are used in healthcare with ethnic minority populations. The corpus of data in training any future AI models must represent the population they are deployed, to mitigate current research trends that show it is unrepresentative of different populations. By addressing these considerations, we can harness the power of AI models to drive positive change in healthcare while promoting fairness and inclusivity.

Declarations

Competing Interests: All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years. KK is the Director for the Real World Evidence Unit and Director for Ethnic Health Research at the University of Leicester.

Funding: MRA, CL and KK are supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC EM) and the NIHR Leicester Biomedical Research Centre (BRC). AW is supported by Health Data Research UK Cambridge, Wellcome Genome Campus, British Heart Foundation Centre of Research Excellence, NIHR Blood and Transplant Research Unit in Donor Health and Genomics, all based at University of Cambridge, Cambridge, UK and British Heart Foundation Data Science Centre, Health Data Research UK, London, UK.

Ethics approval: Not applicable.

Guarantor: MRA is the guarantor. He accepts full responsibility for the work.

Contributorship: All authors (MRA, CL, AW, and KK) have made substantial contributions to the conception, drafting the manuscript, and critically revising it for important intellectual content. Furthermore, all authors have provided final approval of the version to be published.

Provenance: Not commissioned; editorial review.

ORCID iD: Mohammad R Ali https://orcid.org/0000-0002-4765-615X

References

1. Rajpurkar P, Chen E, Banerjee O and Topol EJ. AI in health and medicine. *Nat Med* 2022; 28: 31–38.
2. Radford A, Narasimhan K, Salimans T and Sutskever I. Improving language understanding by generative pre-training. 2018. See https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (last checked 24 March 2023).
3. Haupt CE and Marks M. AI-generated medical advice – GPT and beyond. *JAMA*. 2023; 329: 1349–1350. https://doi.org/10.1001/jama.2023.5321
4. Raunak V, Menezes A and Junczys-Dowmunt M. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. P1172–1183. (Association for Computational Linguistics, 2021).
5. Lee P, Bubeck S and Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233–1239.
6. Pan D, Sze S, Minhas JS, Bangash MN, Pareek N, Divall P, et al. The impact of ethnicity on clinical outcomes in COVID-19: a systematic review. *EClinicalMedicine* 2020; 23: 100404.
7. Simon P. Collecting ethnic statistics in Europe: a review. *Ethnic Racial Studies* 2012; 35: 1366–1391.
8. Bodicoat DH, Routen AC, Willis A, Ekezie W, Gillies C, Lawson C, et al. Promoting inclusion in clinical trials – a rapid review of the literature and recommendations for action. *Trials* 2021; 22: 880.
9. Routen A, Akbari A, Banerjee A, Katikireddi SV, Mathur R, McKee M, et al. Strategies to record and use ethnicity information in routine health data. *Nat Med* 2022; 28: 1338–1342.
10. Dean L. Clopidogrel therapy and CYP2C19 genotype. Medical Genetics Summaries. National Center for Biotechnology Information (US); 2018. See www.

ncbi.nlm.nih.gov/books/NBK84114/ (last checked 8 March 2022).

11. Williams WW, Hogan JW and Ingelfinger JR. Time to eliminate health care disparities in the estimation of kidney function. *N Engl J Med* 2021; 385: 1804–1806.
12. Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004; 364: 937–952.
13. Blicharska M, Smithers RJ, Kuchler M, Agrawal GK, Gutiérrez JM, Hassanali A, et al. Steps to overcome the North–South divide in research relevant to climate change policy and practice. *Nature Clim Change* 2017; 7: 21–27.
14. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Chapter 8: assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd Edition. Chichester, UK: John Wiley & Sons, 2019, p. 398. See www.wiley.com/en-us/Cochrane+Handbook+for+Systematic+Reviews+of+Interventions%2C+2nd+Editi on-p-9781119536628 (last checked 16 May 2023).
15. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011; 64: 383–394.