An ethics assessment tool for artificial intelligence implementation in healthcare: CARE-AI

Yilin Ning[1], Xiaoxuan Liu[2,3], Gary S. Collins[4], Karel G. M. Moons[5], Melissa McCradden[6,7,8], Daniel Shu Wei Ting[1,9], Jasmine Chiat Ling Ong[10], Benjamin Alan Goldstein[11], Siegfried K. Wagner[12,13], Pearse A. Keane[12,13], Eric Topol[14], Nan Liu[1,15,16*]

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

[2]College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[3]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

[4]UK EQUATOR Centre, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[5]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[6]Department of Bioethics, The Hospital for Sick Children, Toronto, Ontario, Canada

[7]Genetics and Genome Biology, Peter Gilgan Centre for Research and Learning, Toronto, Ontario, Canada

[8]Dalla Lana School of Public Health, Toronto, Ontario, Canada

[9]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

[10]Division of Pharmacy, Singapore General Hospital, Singapore, Singapore

[11]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[12]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK

[13]Institute of Ophthalmology, University College London, London, UK

[14]Scripps Research Translational Institute, Scripps Research, La Jolla, CA, USA

[15]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore

[16]Institute of Data Science, National University of Singapore, Singapore, Singapore

*Correspondence: Nan Liu, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore

Email: liu.nan@duke-nus.edu.sg

Artificial intelligence (AI)-powered prediction models in the healthcare domain can lead to ethical concerns on their implementation and upscaling. For example, AI prediction models can hinder clinical decision-making if they advise different diagnoses or treatments by sex or race without clear justification, or they can directly harm patients if they incorrectly guide termination of life-sustaining therapies due to poor predicted prognosis that ultimately realizes the prediction. Recent recommendations, guides (e.g., the WHO guidance on ethics and governance of AI for health and the Dutch guideline on AI for healthcare) and legislation (e.g., the European Union AI Act, and the White House Executive order on Safe, Secure, and Trustworthy Development and Use of AI) have outlined important considerations around key principles for implementation of AI, including ethical considerations.[1,2] Health systems have responded by establishing governance committees and processes to ensure the safe an equitable implementation of AI tools.[3] However, existing recommendations do not explicitly focus on or provide an assessment tool to identify and mitigate ethical issues when implementing AI prediction models in healthcare practice, including the public health domain.

The development and validation of AI prediction models has benefited from detailed reporting and risk of bias tools such as TRIPOD+AI and PROBAST (with its forthcoming AI extension[4]) that highlighted fairness and bias control, and CLAIM[5] for reporting AI medical imaging studies that highlighted data privacy, security and interpretability. However, when planning the implementation of a rigorously developed and well-performing AI prediction model in daily healthcare practice, existing recommendations and guidance on ethics are sparse and lack operational detail. For example, the DECIDE-AI reporting guideline[6] contains a small number of ethics-related recommendations for early clinical evaluation of AI concerning equity, safety and human-AI interaction, and FUTURE-AI[7] provides some recommendations based on six principles (Fairness, Universality, Traceability, Usability, Robustness, and Explainability) in model design, development, validation and deployment. There lacks established and bioethics-centric delivery science toolkit for responsible AI implementation in healthcare.[8]

As highlighted by the United Kingdom National Screening Committee's approach to reviewing evidence on AI in breast cancer screening,[9] much effort is required to translate scientific evaluations of AI to health outcomes with direct relevance to the patient, e.g., to understand the clinical risks and benefits of AI-based diagnoses. A more holistic framework for evaluating digital health technologies and accounting for health quality and equity is summarised in the National Institute for Health and Care Excellence (NICE) Evidence standards framework for digital health technologies. Applying established ethical principles in healthcare practice, we are setting out to develop a new assessment tool that consolidates existing guidance, identifies gaps, and provides recommendations to promote implementation of fair, trustworthy and thus ethically responsible AI prediction models to improve health outcomes – called the Collaborative Assessment for Responsible and Ethical AI Implementation (CARE-AI) tool. In addition to disease diagnosis and prognosis that have been the focus of existing recommendations on prediction models, CARE-AI will include less discussed yet important applications including AI therapeutics that enhance drug discovery and development.[10] While focusing on AI prediction models, CARE-AI will extend to cover prediction models not necessarily involving AI (e.g., regression models) where applicable.

CARE-AI is a tool to guide responsible clinical implementation of AI-based prediction models, primarily aiming to assist healthcare professionals and hospital or other medical care leaderships, but is also useful for other stakeholders including ethical review boards, funding agencies, editorial boards, and regulatory agencies. Specifically, the CARE-AI tool will comprise a list of prompting questions and a decision tree to operationalize recommended practice. We have ensembled a working group of international researchers with diverse backgrounds to develop the CARE-AI tool, including healthcare professionals, bioethicists, data scientists, statisticians and AI researchers, prediction methodologists, editors, and guideline developers. To ensure rigorous development of the CARE-AI tool, we will follow methodology as provided by the EQUATOR Network guidance[11] to reach consensus when evaluating and formalizing each item in the assessment tool with a broader group of researchers and stakeholders (including patient representatives). We will pilot test the assessment tool among healthcare professionals, prediction model developers and other stakeholders to evaluate usability. Complementary to existing and upcoming guidance on the development and evaluation of AI for healthcare such as TRIPOD+AI, PROBAST+AI and other AI guidelines, CARE-AI will provide recommendations to facilitate the translation of trustworthy AI prediction models, and notably the transportability of such tools across all relevant subgroups and end-users including minority groups. An international collaboration is crucial for accommodating the complexity and diversity of healthcare systems worldwide, the heterogeneity of legal foundations and policy guidance, and the ethical challenges they face. We therefore welcome interested global stakeholders to join this collaborative effort.

References:

1. Sounderajah, V. *et al.* Ethics methods are required as part of reporting guidelines for artificial intelligence in healthcare. *Nat. Mach. Intell.* **4**, 316–317 (2022).
2. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit. Med. 2022 51* **5**, 1–13 (2022).
3. Economou-Zavlanos, N. J. *et al.* Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. *J. Am. Med. Inform. Assoc. JAMIA* **31**, 705–713 (2024).
4. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).

5. Tejani, A. S. *et al.* Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for reporting AI research. *Nat. Mach. Intell.* **5**, 950–951 (2023).
6. Vasey, B. *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
7. Lekadir, K. *et al.* FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. Preprint at https://doi.org/10.48550/arXiv.2309.12325 (2023).
8. Li, R. C., Asch, S. M. & Shah, N. H. Developing a delivery science for artificial intelligence in healthcare. *Npj Digit. Med.* **3**, 1–3 (2020).
9. Taylor-Phillips, S. *et al.* UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *Lancet Digit. Health* **4**, e558–e565 (2022).
10. Pun, F. W., Ozerov, I. V. & Zhavoronkov, A. AI-powered therapeutic target discovery. *Trends Pharmacol. Sci.* **44**, 561–572 (2023).
11. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).