frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 12 September 2023
DOI 10.3389/frai.2023.1282800

Editorial: Explainable artificial intelligence for critical healthcare applications

Zhe He[1]*, Rui Zhang[2], Gayo Diallo[3], Zhengxing Huang[4] and Benjamin S. Glicksberg[5]

[1]School of Information, Florida State University, Tallahassee, FL, United States, [2]Division of Computational Health Sciences, University of Minnesota, Minneapolis, MN, United States, [3]Team AHeaD, Bordeaux Population Health INSERM U1219, University of Bordeaux, Bordeaux, France, [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China, [5]Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Thomas Hartung,
Johns Hopkins University, United States

*CORRESPONDENCE
Zhe He
✉ zhe@fsu.edu

RECEIVED 24 August 2023
ACCEPTED 28 August 2023
PUBLISHED 12 September 2023

CITATION
He Z, Zhang R, Diallo G, Huang Z and Glicksberg BS (2023) Editorial: Explainable artificial intelligence for critical healthcare applications. *Front. Artif. Intell.* 6:1282800.
doi: 10.3389/frai.2023.1282800

KEYWORDS

explainable artificial intelligence, healthcare application, health informatics, electronic health records, medical image

Editorial on the Research Topic
Explainable artificial intelligence for critical healthcare applications

Introduction

With the recent advancements of Artificial Intelligence (AI), its applications in human health continue to be of focal interest. With the enormous amount of health-related data being generated and aggregated, especially from electronic health records (EHRs) and medical images, we are faced with an AI revolution that has been shown to accelerate clinical research, optimize healthcare quality, provide data-driven clinical decision support, and ultimately improve patients' health outcomes (Rajpurkar et al., 2022). Further, with the recent explosion of interest in generative large language model-based tools such as ChatGPT, AI is expected to further revolutionize healthcare delivery. However, there are also growing concerns that AI, with inadequate oversight and regulation, may advertently cause harm with its inherent biases, lack of transparency, and lack of interpretability, which are vital to garner its trustworthiness, especially in clinical practice. Although there are ethics-based AI principles and reporting guidelines such as Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis-AI (TRIPOD-AI) (Collins et al., 2021) and MINimum Information for Medical AI Reporting (MINIMAR) (Hernandez-Boussard et al., 2020), there isn't yet any guideline that covers transparency, fairness, explainability, and ethics at the same time. Recently, the US National Academy of Medicine is working with leading bioethics, research, and patient advocacy organizations to develop an AI Code of Conduct to support the equitable and responsible use of AI for health care and research. Similarly, as part of its digital strategy, the European Commission wants to regulate AI (the AI Act) to ensure better conditions for the development and use of this innovative technology (European Commission, 2023). Health and biomedical informatics are at the forefront of medical AI because informatics researchers study the optimal use of medical knowledge-driven techniques to benefit clinicians, clinical researchers, and patients, and without adequate buy-in, medical AI will not realize its intended value.

Why is explainability a major concern in medical AI and crucial for its trustworthiness? Traditional machine learning highly relies on laborious feature engineering, which may not be adequate for discovering complex patterns in high-dimensional EHRs or medical images, resulting in suboptimal performance. In contrast, deep learning (DL) methods enable machines to automatically detect the intricate relationships among features and extract meaningful knowledge from data. More recently, advanced DL models such as transformers and auto-encoders have shown promise in representing complex clinical data from multiple modalities, leading to significantly improvements in downstream tasks such as patient similarity identification, disease onset prediction, and deep phenotyping. However, DL algorithms are often perceived as black-box models because they incorporate high-degree interactions between input features through a multi-layer nonlinear structure with numerous neurons. To ensure the trustworthiness of DL models and justify their predictions in high-stakes healthcare applications, it is crucial for medical professionals to understand how the system generates specific predictions. Such explanations play an essential role in ensuring fairness and accountability in the clinical decision-making process. This aspect is important, as a single incorrect prediction from the system can have real-world consequences that may lead to serious medical errors (Obermeyer et al., 2019). Recognizing the significance of transparency, the European Union's General Data Protection Regulation (GDPR) was recently enacted to require organizations that use patient data for classifications and recommendations to provide on-demand explanations. Additionally, the White House has developed guidance on AI applications that emphasizes transparency as a fundamental principle for the stewardship of AI applications. Sufficient explanations of the AI models allow medical doctors to comprehend and trust AI-based clinical decision support systems. Consequently, research on explainable AI (XAI) in healthcare is increasing to address these concerns (Payrovnaziri et al., 2020).

To establish an open-access platform for researchers to disseminate novel methods and applications of XAI in critical healthcare applications, we dedicate this Research Topic to include original research articles that present innovative XAI methodologies and applications in healthcare that aim to ensure fairness, accountability, and trustworthiness of AI systems. Following a rigorous review, revision, and selection process, we finally accepted four articles including three original research articles and one review article.

Published studies in this Research Topic

In Kiseleva et al., the authors from Belgium and France reviewed the legislation and policies regarding transparency in health AI models in the European Union and suggested that transparency is an umbrella concept that includes interpretability and explainability and it should be achieved through both technical and non-technical measures. The article also discusses the importance of making AI in healthcare externally transparent to patients such that patients should be informed and give informed consent to use it. They also stressed that the implementation of a transparency system must be contextualized.

In An et al., a team of researchers from the US addressed a critical problem in AI-based breast cancer screening, the lack of high-quality pixel-level annotated data for anomalous tissue by introducing LatentCADx, a DL segmentation model that can precisely annotate cancer lesions underlying hand-drawn annotations. This architecture is based on the ResNet convolution and allows weights learned from a ResNet model to be transferred to another ResNet model as well as selective training of layers. The performance of LatentCADx using a publicly available dataset of 2,620 mammogram case files reached 0.97 AUROC. In addition, qualitative evaluation of the segmentation showed great clarity and specificity with weakly-supervised segmentation, which ensures the interpretability of the model.

In Fujihara et al., a research team from Japan developed a ML model to predict changes in future weight in the subsequent 3 years using 67,021 Japanese individuals who had initial health examinations at the Niigata Association of Occupational Health. The predictors include body mass index, smoking and alcohol intake status, exercise habit, skipping breakfast, walking speed, eating speed, and other medical conditions such as history of dyslipidemia, diabetes, and hypertension. They used heterogeneous mixture learning technology (HMLT) which allows a mixture of data following different patterns. This technique uses predicted weights after 1 year to predict weights after 2 years, and then 3 years. The overall root mean square error (RMSE) was 1.914. For interpretability, HMLT can generate a decision-tree-like structure with five linear models based on rules that maximize the prediction accuracy. As expected, they found people with healthy lifestyle habits tended to have weight reduction, while people with unfavorable lifestyle habits tended to increase weight.

In Chu et al., researchers developed ML models to predict physical functions after older adults are discharged from hospitalization in Taiwan. In the experiments, 52 clinical features were considered when evaluating XGBoost, random forest, logistic regression models, which yield AUROC of 98%, 98%, and 97%, respectively. Using SHAP values for interpretability assessment, they found that for the XGBoost model, activity of daily living at baseline, activity of daily living at admission, and mini nutritional status during admission were the top three predictors.

To summarize, XAI is instrumental in improving potential buy-in among healthcare professionals. But XAI alone is not sufficient for the trustworthiness of AI, other factors including fairness of the model, transparency of the biases in the training data, external validity of the models, and clinical utility of the findings are all important for the equitable and responsible use of AI for healthcare. As AI models make their way into the clinic via Software as a Medical Device (SaMD) designation, we are looking forward to seeing authoritative ethical AI guidelines and more exciting research that pushes this research agenda forward.

Author contributions

ZHe: Writing—original draft, Writing—review and editing. RZ: Writing—review and editing. GD: Writing—review and editing. ZHu: Writing—review and editing. BG: Writing—review and editing.

Funding

ZHe was funded in part by the National Institutes of Health (NIH), National Library of Medicine grant 5R21LM013911-02, NIH National Institute on Aging grants 3R21AG061431-02S1, 5R21AG061431-02, and 5R01AG064529-05, NIH National Institute on Alcohol Abuse and Alcoholism grant 5P01AA029547-02, and US Environmental Protection Agency grant 84063201. RZ was funded in part by the National Center for Complementary and Integrative Health (NCCIH) under grant number 2R01AT009457 and NIA under grant number 1R01AG078154, and National Cancer Institute under grant number 1R01CA287413.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Collins, G. S., Dhiman, P., Navarro, C. L. A., Ma, J., Hooft, L., Reitsma, J. B., et al. (2021). 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. doi: 10.1136/bmjopen-2020-048008

European Commission (2023). *An European Approach to Artificial Intelligence | Shaping Europe's digital future* [Internet]. Available online at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed August 24, 2023).

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., and Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inform. Assoc.* 27, 2011–2015. doi: 10.1093/jamia/ocaa088

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., et al. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J. Am. Med. Inform. Assoc.* 27, 1173. doi: 10.1093/jamia/ocaa053

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nat Med.* 28, 31–38. doi: 10.1038/s41591-021-01614-0