False Conflict and False Confirmation Errors Are Crucial Components of AI Accuracy in Medical Decision Making

Rosenbacke, Rikard; Melhus, Åsa; Stuckler, David

Document Version
Final published version

Published in:
Nature Communications

DOI:
10.1038/s41467-024-50952-3

Publication date:
2024

nature communications

Matters arising

https://doi.org/10.1038/s41467-024-50952-3

False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making

Received: 9 April 2024

Accepted: 25 July 2024

Published online: 13 August 2024

Check for updates

Rikard Rosenbacke[1] ✉, Åsa Melhus[2] & David Stuckler[3]

ARISING FROM T. Chanda et al. *Nature Communications* https://doi.org/10.1038/s41467-023-43095-4 (2024)

We welcome the recent study from January 2024 by Chanda and colleagues[1] in *Nature Communications*, as a substantial advance on integrating Explainable Artificial Intelligence (XAI) into dermatological practice. Importantly, it shows that AI can enhance diagnostic accuracy, trust, and confidence among dermatologists.

When physicians make decisions with AI, three types of errors can occur (Table 1): (i) false confirmation error—when the physician and AI agree but both are wrong; (ii) false conflict error—when the physician is correct, AI is incorrect, and the physician change diagnosis; and (iii) true conflict error—when the physician is incorrect but AI is correct, and the physician override the correct AI diagnosis.

In their paper, Chanda and colleagues consider only overall accuracy, which masks key decision-making threats and overlook the specific user groups that stand to gain the most from AI applications.

We revisited their published data, quantifying these errors (albeit we note that without full access to their original data, we cannot make precise calculations). With a mean AI error rate of 19.6%, combined with a mean clinician error rate of 33.8%, the likelihood of both being inaccurate, or a false confirmation, is 6.6%. Applying these calculations to the worse performing clinicals (lowest quartile mean accuracy 50.3%) increases the false confirmation rate to 9.7%.

We also found evidence consistent with false conflict errors for high-performing physicians. A sub-analysis of the best-performing physicians reveals that their performance deteriorates with AI support. A sub-analysis of the 15 best-performing clinicians matched or exceeded AI accuracy (80.4%), with their initial accuracy averaging 87.3% but dipping to 77.1% once AI was introduced in phase 2 and 81.5% with XAI in phase 3, possibly due to errors from relying on AI when it falsely conflicted with their own correct diagnosis. Trust in AI is not, by definition, better since it increases false conflict errors for the best performers.

Finally, we found that AI, for the lowest-performing clinicians, helped stamp out true conflict errors. For the lowest-performing quartile of clinicians studied by Chanda and colleagues, accuracy improved from 50.3% to 66.6% and 65.9%, respectively, during the three phases of their study. In theory, if these low-performing physicians fully trusted AI, their accuracy could have risen at least to 80.4% by simply eliminating true conflict errors.

Recent studies are beginning to delve deeper into how physicians respond to conflicts with AI. The most common and discussed error occurs when physicians tend to override a correct AI diagnosis in cases of true conflict error. Previous studies found that this arises from distrust in the AI's "black box" logic[2–5]. In cases of false conflict errors, however, the physicians tended to express doubt and over-rely upon AI, especially when uncertain about their initial diagnosis. When explanations are added to the AI diagnoses (as XAI), it tends to mitigate true conflict errors but exacerbate false conflict errors. This phenomenon whereby even mere exposure to explanations can induce over-reliance on AI has been documented in several studies[6–9]. Finally, false confirmation is perhaps the most pernicious; it reinforces trust in AI, while perpetuating clinical errors. These false confirmation errors remind us of the confirmation bias highlighted by Ghassemi and colleagues[10]. This issue is likely present in the study conducted by Chanda and colleagues, though it was not explicitly addressed.

Given that explainable AI can assist physicians in determining whether AI diagnoses align with evidence-based medicine and that explanations are essential for meeting the trustworthiness and transparency requirements of the EU AI Act 2024, it still potentially introduces new sources of errors, such as false conflict and false confirmation errors. One intervention could be introducing more complex diagnoses instead of simple yes/no decisions. Another promising technique is conformal predictions, which shifts the focus of

Table 1 | Potential sources of error in human-AI/XAI collaboration

	Physician right	Physician wrong
AI right	Correct	True conflict error
AI wrong	False conflict error	False confirmation error

[1]Centre for Corporate Governance, Department of Accounting, Copenhagen Business School, Copenhagen, Denmark. [2]Department of Medical Sciences/Section of Clinical Microbiology, Uppsala University, Uppsala, Sweden. [3]Department of Social and Political Science, Bocconi University, Milano, Italy. ✉e-mail: rr.ccg@cbs.dk; rikard@rosenbacke.com

the AI model from pinpointing a single accurate clinical recommendation to providing the clinician with a range of possibilities tailored to the individual patient, allowing for further investigation[11]. Further research is needed to better understand interventions to avoid new human-AI collaboration errors.

We believe a more precise identification of these errors and in whom they occur creates tremendous potential to tap the full potential and promise of AI-supported decision-making.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

References

1. Chanda, T. et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nat. Commun.* https://doi.org/10.1038/s41467-023-43095-4 (2024).
2. Gaube, S. et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci. Rep.* **13**, 1383 (2023).
3. Kumar, A., Manikandan, R., Kose, U., Gupta, D. & Satapathy, S. C. Doctor's dilemma: evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Trans. Multimed. Comput. Commun. Appl.* **17**, 1–26 (2021).
4. You, S., Yang, C. L. & Li, X. Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation? *J. Manag. Inf. Syst.* **39**, 336–365 (2022).
5. Martínez-Agüero, S. et al. Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. *Futur. Gener. Comput. Syst.* **133**, 68–83 (2022).
6. Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. *Int. J. Hum. Comput. Stud.* **169**, 102941 (2023).
7. Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. Explainable recommendation: when design meets trust calibration. *World Wide Web* **24**, 1857–1884 (2021).
8. Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N. & Ali, R. Nudging through friction: an approach for calibrating trust in explainable AI. In *Proceedings of 2021 8th IEEE International Conference on Behavioural and Social Computing, BESC 2021* (IEEE, 2021).
9. Kliegr, T., Bahník, Š. & Fürnkranz, J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif. Intell.* **295**, 103458 (2021).
10. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
11. Banerji, C. R. S., Chakraborti, T., Harbron, C. & MacArthur, B. D. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nat. Med.* **29**, 2996–2998 (2023).

Author contributions

R.R., the paper's main author, developed the initial research concept. All authors (R.R., Å.M., and D.S.) contributed to the refinement of the research idea, the analysis and interpretation of data, and have undertaken critical revisions of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-50952-3.

Correspondence and requests for materials should be addressed to Rikard Rosenbacke.

Peer review information *Nature Communications* thanks Tapabrata Chakraborty and Nianyin Zeng for their contribution to the peer review of this work.