npj | digital medicine

Comment

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01255-w

Ethical guidance for reporting and evaluating claims of AI outperforming human doctors

Jojanneke Drogt, Megan Milota, Anne van den Brink & Karin Jongsma

Claims of AI outperforming medical practitioners are under scrutiny, as the evidence supporting many of these claims is not convincing or transparently reported. These claims often lack specificity, contextualization, and empirical grounding. In this comment, we offer constructive ethical guidance that can benefit authors, journal editors, and peer reviewers when reporting and evaluating findings in studies comparing AI to physician performance. The guidance provided here forms an essential addition to current reporting guidelines for healthcare studies using machine learning.

An increasing number of academic reports contend that Artificial Intelligence (AI), in particular machine learning systems, surpasses medical practitioners' performance in various clinical tasks and specialisms[1–3]. These outperformance claims—as we will refer to them in this article—vary in their formulation. Commonly used terms include 'outperform,' 'surpass,' 'exceed,' 'better than', and 'superior to,' but all claims are based on the fundamental assumption that AI can be directly compared to and exceed the expertise of medical practitioners on some level (for examples, see Table 1). These reports have contributed to excitement about AI's potential value for medical contexts and have raised hopes about automating specific diagnostic tasks[4,5]. The claims about outperformance have also been met with skepticism because of methodological flaws in AI studies. For example, it remains uncertain whether AI can outperform medical practitioners in clinical practice because model performance is often evaluated in unrealistic settings[4]. Furthermore, many studies fail to transparently report the circumstances under which AI is compared to medical practitioners, making it hard to verify claims of outperformance[1]. Several scholars have concluded that reports on AI's performance in medicine are "exaggerated,"[6,7] and that it is "time to reality check" these kinds of claims to distinguish genuine potential from hype[4]. Some scholars even warn against using terms like 'outperform' because overpromising language risks being misinterpreted by the media and the public[1,6–8] and may result in "sidestepping ethical concerns leaving no space for issues and criticism" on AI's functioning in medical practice[8].

While the concerns and criticism about outperformance claims are important, there is a dire need for constructive and practically applicable ethical guidance[9] stipulating how outperformance should be reported in studies comparing AI to physician performance. To avoid further misinterpretation and misrepresentation of how well AI performs, it is essential to clarify the components of a well-formulated outperformance claim and the conditions under which studies speak of medical practitioners' 'outperformance' by medical AI. In this comment, we will provide constructive ethical guidance for formulating outperformance claims in the medical domain. We maintain that the guidance we provide is also a necessary addition to reporting guidelines for healthcare studies using machine learning.

Outperformance claims should be specific, contextualized, and empirically grounded

Several studies describe outperformance claims ambiguously and exaggerate or misuse the underlying results. Dhiman et al.[10] have, for example, observed "the inappropriate and unjustified use of strong and leading words to interpret study findings with [the] use of words, such as 'superior' and 'outperforms,' when comparing model performance." This is problematic because such overinterpretation of study findings (or 'spin') can influence readers' judgments, and findings may not translate to clinical practice as well as suggested, posing harm to patients when AI is implemented on the basis of overly positive claims[10]. We, therefore, urge authors, journal editors, and peer reviewers to check whether outperformance claims are specific, contextualized, and empirically grounded. In the next sub-sections, we elaborate on these three recommendations and provide examples of how outperformance claims should (not) be formulated.

Specification. One problem with AI outperformance claims is their insufficient specification of what sort of AI systems and human practitioners have been compared. For instance, some articles simply refer to 'the AI,' 'the algorithm,' or 'the model,' while others leave the medical practitioner unspecified when formulating an outperformance claim. We also observed formulations like 'physicians were outperformed by AI' in several articles[11,12]. Even when authors specify these aspects in other parts of the article, such unspecific formulations of outperformance claims can be interpreted as meaning that AI generally outperforms (expert) physicians. This conclusion is an overstatement; in these studies, only some medical practitioners were outperformed by an algorithm in very specific tasks and under highly controlled circumstances. As Nagendran et al.[7] also warn, hiding such information can be harmful as it may prohibit readers from gaining a full understanding of the claim that is made and encourage conclusions broader than the evidence allows[7]. This may be especially true for readers who do not have the tools or time for an in-depth reading of an academic article.

Contextualization. Aside from specificity, outperformance claims should also be adequately contextualized. Transparency about the context, circumstances, and limitations of AI's performance can help avoid misinterpretations such as 'AI outperforms physicians; *therefore*, it can do

Table 1 | Examples of outperformance claims

Bian Y, Zheng Z, Fang X, Jiang H, Zhu M, Yu J, et al. (2022)	"An artificial intelligence model outperformed radiologists and clinical and radiomics models for prediction of lymph node metastasis at CT in patients with pancreatic ductal adenocarcinoma."[21]
Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. (2019)	"The performance of the convolutional neural network exceeded that of orthopedic surgeons in detecting intertrochanteric hip fractures from proximal femoral radiographs under limited conditions."[22]
Iwaki T, Akiyama Y, Nosato H, Kinjo M, Niimi A, Taguchi S, et al. (2023)	"We constructed the first DL system that recognizes HLs with accuracy exceeding that of humans."[23]
Ding L, Liu G-W, Zhao B-C, Zhou Y-P, Li S, Zhang Z-D, et al. (2019)	"Faster R-CNN surpasses radiologists in the evaluation of pelvic metastatic LNs of rectal cancer, but is not on par with pathologists."[24]
Kaddoura T, Vadlamudi K, Kumar S, Bobhate P, Guo L, Jain S, et al. (2016)	"We developed an automated speech-recognition-inspired classification algorithm for the acoustic diagnosis of PH that outperforms physicians that could be used to screen for PH and encourage earlier specialist referral."[25]

Table 2 | Examples of conclusions drawn on the basis of outperformance claims

Hung J-Y, Chen K-W, Perera C, Chiu H-K, Hsu C-R, Myung D, et al. (2022)	"The AI model showed better performance than the non-ophthalmic physician group in identifying referable blepharoptosis (…) correctly. *Therefore*, artificial intelligence aided tools have the potential to assist in the diagnosis and referral of blepharoptosis for general practitioners."[26] [italics added by authors]
Nishida N, Yamakawa M, Shiina T, Mekada Y, Nishida M, Sakamoto N, et al. (2022)	"The performance of the AI models surpasses that of human experts in the four-class discrimination and benign and malignant discrimination of liver tumors. *Thus*, the AI models can help prevent human errors in US diagnosis."[27] [italics added by authors]

everything a physician can do.' In our analysis of the existing literature, we found research articles that failed to mention what AI can and cannot do in practice. Other authors discussed what they believe to be the practical take-away(s) from an outperformance claim for medical practice but remained vague on how they arrived at their conclusion (for examples, see Table 2). In many of the examples, authors seemed to assume that if AI performs better than physicians, it can play a meaningful role in their diagnostic process. However, it's not apparent from (out)performance alone which tasks AI can support and how it can benefit specific medical practices. If a research study is conducted well, an outperformance claim indicates the algorithm is sound in a specific context. However, it does not necessarily mean the algorithm offers the best or most feasible solution to current problems in medical practice.

Empirical grounding. Finally, an outperformance claim should be robustly supported by evidence rather than speculation. Without empirical evidence, outperformance claims risk losing connection to reality and distracting from the current possibilities of AI. We encountered several speculative formulations of outperformance claims in the articles we analyzed; some (1) stretched the results of the study, (2) made a direct hypothesis about the future following the current status quo, or (3) formulated outperformance as an inevitable result of current progress (Table 3). An outperformance claim thereby becomes framed as a speculative belief, rather than a methodologically supported statement. While such a belief may play a distinct role in advancing AI systems and can stimulate developers to further strive for outperformance, it should not be confused with data-supported statements about AI's current capabilities.

From guidance to practice

The recommendations above can guide the formulation of outperformance claims and thus diminish confusion and misinterpretation of what AI can do in medical practice. Fortunately, nuanced and thoughtfully formulated claims that avoid the above-mentioned pitfalls exist as well. Kong et al.[13] formulate a largely specific and non-speculative outperformance claim: "our developed model achieved a prediction accuracy of 90.7% on the internal test dataset and outperformed the performance of ten junior internal medicine physicians (89.0%)"[13]. By including the percentages as well as the number and kind of medical practitioners in the claim, the authors specify and, therefore, clarify what was found in their study. Their attempt at transparent reporting facilitates a critical appraisal of their findings: there is a slight difference in accuracy, and they compared the model with *junior* internal medicine physicians. Does this mean the model is also better than internal medicine physicians? Greater transparency thus supports understanding the study's findings and reflecting on the implications. Some authors also contribute to the reader's understanding of the findings by directly providing nuances; this is illustrated by Angkurawaranon et al.[14], who state "our results demonstrate that the high level of accuracy achieved by the deep learning model, (0.89), outperforms the residents with regard to sensitivity (0.82) but still lacks behind in specificity (0.90)"[14], which is a highly transparent account of their findings and helps readers to see that AI can outperform medical practitioners in a single area.

Discussion: a need for transparent reporting

Many newly developed reporting guidelines, such as TRIPOD-AI[15], STARD-AI[16], and SPIRIT-AI[17], emphasize the urgent need for more transparent reporting in healthcare studies using machine learning systems. These guidelines also provide recommendations for stakeholders such as researchers, journal editors, peer reviewers, and the public on evaluating AI studies and their results[15].

Some of the items on those checklists are similar to the guidance we have formulated. TRIPOD-AI, for instance, recommends focusing on the usability of the model in the context of current care when reporting on the results, which links to the point we make about contextualization[15]. SPIRIT-AI suggests that the type of AI and its intended use should be specified in the title, which connects to our point on specification[17]. Nevertheless, specific items on how claims on AI performance should be formulated, seem to be absent in these checklists. This is problematic, as it still leaves the door open for unjustified or misleading outperformance claims. The guidance we have provided is thereby an essential addition to current guidelines for transparent reporting.

An overarching problem for transparent reporting is that adherence to and citing of reporting guidelines is poor among AI versus physician

Table 3 | Speculative outperformance claims

Stretching the results of the study	"Machine learning applied to nasal air pressure tracings is feasible and *may exceed* the diagnostic performance of expert clinicians."[28] [italics added by authors]
	"Fully automated BPE assessments for breast MRIs *could be more accurate* than BPE assessments from radiology reports."[29] [italics added by authors]
A direct hypothesis on the future following the current status quo	"The CNN (…) can now provide better results than doctors. In the future, as training data evolves and improves, we *anticipate* that AI will perform significantly better than physicians."[30] [italics added by authors]
	"From the study, it is concluded that when the patient base reaches a certain number, the diagnostic accuracy of the machine-assisted diagnosis system *will exceed* the doctor's expertise."[31] [italics added by authors]
Outperformance seen as an inevitable result of current progress	"It seems inevitable that diagnostic and recommender artificial intelligence models *will ultimately* reach a point when they outperform human clinicians"[32] [italics added by authors]
	"Someday, perhaps soon, diagnostics generated by machine learning (ML) *will have demonstrably better success* than those generated by human doctors"[33] [italics added by authors]

performance studies[3,7]. To stimulate more transparent reporting, creating a more unified standard that refers to relevant recommendations for reporting on different stages of AI development and enforcing it may be necessary to consolidate current guidelines and ethical guidance. While many guidelines are still rather new and might benefit from increased visibility, as is for instance recognized and stimulated by the EQUATOR network[18], another possible reason for the lack of adherence to guidelines could be the presence of numerous recommendations, leading to uncertainty on which checklists authors and journals should rely. A more unified and harmonized standard may also be helpful, as a combination of guidelines may often be appropriate[19], and it enables bundling general guidance for dealing with common problems in reporting on AI[20].

Conclusion

Claims of AI outperforming medical practitioners are under scrutiny, as the evidence supporting many of these claims is not convincing or transparently reported. This is concerning, as it indicates that the term 'outperformance' is misappropriated in the literature and can contribute to misleading perceptions about medical AI's current performance. In addition to current reporting guidelines for transparent reporting in healthcare studies using machine learning applications, constructive and practically applicable ethical guidance for how such claims should be posed is urgently needed to avoid misinterpretation and misrepresentation of how AI performs; this is why we have made three recommendations in this commentary. Specifically, we maintain that an outperformance claim should be precisely formulated, accompanied by a clear description of its implications, and devoid of speculation. We believe these steps can support creating greater transparency of conditions under which outperformance claims are made and evaluating AI's genuine potential for supporting medical practitioners.

Jojanneke Drogt ✉, Megan Milota, Anne van den Brink & Karin Jongsma
University Medical Center, Utrecht, The Netherlands.
✉e-mail: j.m.t.drogt@umcutrecht.nl

Received: 27 May 2024; Accepted: 8 September 2024;
Published online: 02 October 2024

References

1. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health* **1**, e271–e297 (2019).
2. Lebovitz, S., Levina, N. & Lifshitz-Assaf, H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Q.* **45**, 1501–1525 (2021).
3. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digital Health* **6**, e367–e373 (2024).
4. Wilkinson, J. et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digital Health* **2**, e677–e680 (2020).
5. Fogel, A. L. & Kvedar, J. C. Artificial intelligence powers digital medicine. *NPJ Digital Med.* **1**, 5 (2018).
6. BMJ. *Concerns over 'exaggerated' study claims of AI outperforming doctors: Misleading claims fuel hype and pose a patient safety risk, warn researchers*, www.sciencedaily.com/releases/2020/03/200325212159.htm (2020).
7. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *bmj* **368**, m689 (2020).
8. Bunz, M. & Braghieri, M. The AI doctor will see you now: assessing the framing of AI in news coverage. *AI Society* **37**, 9–22 (2022).
9. Morley, J. et al. Operationalising AI ethics: barriers, enablers and next steps. *AI Society* **38**, 411–423 (2023).
10. Dhiman, P. et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J. Clin. Epidemiol.* **157**, 120–133 (2023).
11. Gasulla, Ó. et al. Enhancing physicians' radiology diagnostics of COVID-19's effects on lung health by leveraging artificial intelligence. *Front. Bioeng. Biotechnol.* **11**, 1010679 (2023).
12. Dorr, F. et al. COVID-19 pneumonia accurately detected on chest radiographs with artificial intelligence. *Intell.-Based Med.* **3-4**, 100014 (2020).
13. Kong, Y. et al. Constructing an automatic diagnosis and severity-classification model for acromegaly using facial photographs by deep learning. *J. Hematol. Oncol.* **13**, 88 (2020).
14. Angkurawaranon, S. et al. A comparison of performance between a deep learning model with residents for localization and classification of intracranial hemorrhage. *Sci. Rep.* **13**, 9975 (2023).
15. Collins, G. S. et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj* **385**, e078378 (2024).
16. Sounderajah, V. et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ open* **11**, e047709 (2021).
17. Rivera, S. C. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digital Health* **2**, e549–e560 (2020).
18. EQUATOR network. *Enhancing the QUAlity and Transparency Of health Research*, https://www.equator-network.org (2024).
19. Klontzas, M. E., Gatti, A. A., Tejani, A. S. & Kahn, C. E. Jr AI reporting guidelines: how to select the best one for your research. *Radiology: Artif. Intell.* **5**, e230055 (2023).
20. Flanagin, A. et al. Reporting use of AI in research and scholarly publication—JAMA Network Guidance. *JAMA* **331**, 1096–1098 (2024).
21. Bian, Y. et al. Artificial Intelligence to Predict Lymph Node Metastasis at CT in Pancreatic Ductal Adenocarcinoma. *Radiology* **306**, 160–169 (2022).
22. Urakawa, T. et al. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skelet. Radiol.* **48**, 239–244 (2019).
23. Iwaki, T. et al. Deep Learning Models for Cystoscopic Recognition of Hunner Lesion in Interstitial Cystitis. *Eur. Urol. Open Sci.* **49**, 44–50 (2023).
24. Ding, L. et al. Artificial intelligence system of faster region-based convolutional neural network surpassing senior radiologists in evaluation of metastatic lymph nodes of rectal cancer. *Chin. Med. J.* **132**, 379–387 (2019).
25. Kaddoura, T. et al. Acoustic diagnosis of pulmonary hypertension: automated speech-recognition-inspired classification algorithm outperforms physicians. *Sci. Rep.* **6**, 33182 (2016).
26. Hung, J.-Y. et al. An outperforming artificial intelligence model to identify referable blepharoptosis for general practitioners. *J. Personalized Med.* **12**, 283 (2022).
27. Nishida, N. et al. Artificial intelligence (AI) models for the ultrasonographic diagnosis of liver tumors and comparison of diagnostic accuracies between AI and human experts. *J. Gastroenterol.* **57**, 309–321 (2022).
28. Crowson, M. G. et al. Paediatric sleep apnea event prediction using nasal air pressure and machine learning. *J. Sleep. Res.* **32**, e13851 (2023).

29. Eskreis-Winkler, S. et al. Breast MRI Background Parenchymal Enhancement Categorization Using Deep Learning: Outperforming the Radiologist. *J. Magn. Reson. Imaging* **56**, 1068–1076 (2022).
30. Soydan, Z. et al. An AI based classifier model for lateral pillar classification of Legg–Calve–Perthes. *Sci. Rep.* **13**, 6870 (2023).
31. Zhang, J., Chen, Z., Wu, J. & Liu, K. An intelligent decision-making support system for the detection and staging of prostate cancer in developing countries. *Computational Math. Methods Med.* **2020**, 5363549 (2020).
32. Banja, J. D., Hollstein, R. D. & Bruno, M. A. When Artificial Intelligence Models Surpass Physician Performance: Medical Malpractice Liability in an Era of Advanced Artificial Intelligence. *J. Am. Coll. Radiol.* **19**, 816–820 (2022).
33. Froomkin, A. M., Kerr, I. & Pineau, J. When AIs outperform doctors: confronting the challenges of a tort-induced over-reliance on machine learning. *Ariz. L. Rev.* **61**, 33 (2019).

Acknowledgements
This study was funded by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO)—project number 406.DI.19.089. We want to thank our fellow RAIDIO project members—Sally Wyatt, Flora Lysen, and Shoko Vos—and the members of the AI ethics seminars at the UMC Utrecht for their insightful thoughts and suggestions, which helped to improve the manuscript.

Author contributions
J.D., K.R., and M.M. were responsible for conceptualizing the initial concept for this manuscript. A.B. and J.D. analyzed the literature; J.D. was responsible for drafting and revising the manuscript. K.R., M.M., and A.B. provided critical intellectual input and revisions. All authors approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Jojanneke Drogt.