npj | digital medicine

News & Views

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01344-w

The path forward for large language models in medicine is open

Lars Riedemann, Maxime Labonne & Stephen Gilbert

Large language models (LLMs) are increasingly applied in medical documentation and have been proposed for clinical decision support. We argue that the future for LLMs in medicine must be based on transparent and controllable open-source models. Openness enables medical tool developers to control the safety and quality of underlying AI models, while also allowing healthcare professionals to hold these models accountable. For these reasons, the future is open.

Large language models (LLMs) promise to add significant value to medicine by reducing administrative burden, e.g., through text summarization tasks, improving information retrieval, and enhancing diagnostic and therapeutic procedures[1]. Although there are both skeptics and enthusiasts regarding their likely short- and long-term impact, an increasing adoption can be observed, in particular in medical documentation and summarization[2]. Selecting a LLM and its provider is a challenging strategic decision for early adopters in healthcare, who would like to integrate LLMs into their clinical processes. This decision is not merely technical but strategic, as it reflects broader organizational values and priorities. In this context, a key question arises: to what extent does openness become a decisive factor in choosing an underlying LLM framework for medical applications? In certain healthcare applications and deployment settings, there may be a particular need to prioritize ease of deployment and maintenance and technical reliability, aligning more closely with closed models that offer proprietary support and defined frameworks. Conversely, in the development of locally optimized clinical decision support systems, there may be the need to lean towards open-source LLMs, offering a maximum of transparency, adaptability to the latest medical guidelines and local protocols and for the calibration and optimization for specific medical scenarios, and personal preferences. Here we discuss the advantages and disadvantages of LLMs across the entire openness spectrum, from open-source LLMs, built with the intention to provide full transparency and access to each part, to fully closed models, whose manufacturers intentionally restrict access to the source code, model weights, and training procedures and data. We address high-level approaches to the application of LLMs in medicine, rather than addressing in detail which medical tasks these tools are best suited for, which we and others have addressed elsewhere[3–5].

Why is the LLM openness question so critical for medicine?

Artificial Intelligence (AI) systems are often perceived as black boxes, and parts of the inner technical workings of LLMs are indeed poorly understood by laymen and AI researchers alike. There is a lack of full understanding of these systems, and unintentional subtle changes to such complex systems can lead to a tremendous impact on their output quality. In order to reach a better understanding and allow for continuous improvements, we argue that LLMs deployed in medicine must provide maximum transparency, comprehensive oversight, and maximum control. To operate LLMs reliably as part of a medical device, a deep understanding and control of the LLM architecture and provenance are mandatory. This includes knowledge and/or control over training datasets, methodologies, and update processes[6]. Proprietary closed-source LLMs are in stark contrast to these necessities, as the inner workings and revision processes remain hidden from public scrutiny. Current market-leading, closed-source models do not provide developers of medical devices with either a static model with long-term support, or a model that has a detailed explanation and justification of its changes. In other words, constraint and adaptation processes continue on live models through poorly defined processes that are not transparent to developers. In December 2023, ChatGPT users noticed that the chatbot had become "lazy"[7]. The exact reason for this could neither be identified by researchers nor was it disclosed by its operating company. Furthermore, AI researchers have evaluated the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks and showed that the behavior of the "same" LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs[8]. Furthermore, if the operating company discontinues support for a closed-source LLM, it could have severe consequences for any medical device reliant on it. Current market-leading, closed-source models are regularly updated, and their operators do not provide a persistent stored historical model that can later be interrogated by medical device developers to carry out legally required analysis of the root cause of safety or performance issues identified by themselves or reported by users[6]. Open-source approaches can overcome many of these limitations by maximizing transparency, and by providing medical device developers maximum flexibility to reduce the black-box properties of such systems. They can make sure their medical devices use only a fixed and controlled model and make sure that the model is archived after it is no longer in use so that safety issues can later be investigated if safety reports are received. Furthermore, permissive open-source licenses under which these models can be operated allow further development and adjustments to LLMs as well as free use of the software including for commercial use. From a technical standpoint, LLMs along the full openness spectrum could be integrated into AI systems. However, open-sourced models have clear advantages in the medical domain based on their transparency and enablement of developer control.

How do we discern open from closed-source LLMs?

How can we systematically analyze an LLM with respect to its openness, and how do open-source LLMs compare in terms of performance

to their closed-source counterparts? The question of what counts as open-source in generative AI (GenAI) is poised to take on great importance in light of the EU AI Act that regulates open-source differently from closed-source models, creating an urgent need for practical frameworks for assessing the openness of LLMs. The EU AI Act categorizes LLMs and GenAI foundation models as general-purpose AI (GPAI) models, and splits these into multiple complexly interlinked subclassifications with different requirements depending on the properties of each, which are sometimes narrowly and sometimes broadly defined. An exemption to demonstrating certain aspects of model transparency is enjoyed by GPAI models released under free and open-source licenses if their parameters, including their weights, the information on the model architecture, and the information on model usage, are made publicly available. There are numerous other AI Act requirements that apply to all GPAI models irrespective of whether they are open or closed, and some exceptions only for open-source models not meeting the threshold of models of systemic risk (the definition of which is addressed later in this article). If this comes across to the reader as an impossible labyrinth, that is because it is, and hopefully, later guidance and amendments of laws (which is acknowledged as being necessary in the AI Act) will provide some form of navigability.

Several groups have recently proposed model openness frameworks, taking into consideration the main LLM building blocks[9,10]. In Fig. 1, we provide an exemplary overview of the openness of three popular LLMs across 14 dimensions, based on a framework recently proposed by Liesenfeld et al.[9]. Based on this framework, ChatGPT, powered either by the original GPT-3.5[11] or the more recent GPT-4o[12] model, serves as a prominent example of a closed-source system. Mistral AI's Mistral 7B Instruct[13] model is more open, in particular, because its model weights are publicly available, yet it falls short of full open-source status, in particular, because its manufacturer does not share detailed information of the underlying training data and system architecture. OLMo 7B Instruct[14] produced by the Allen Institute for AI is currently one of the most open models. A granular analysis of all model components demonstrates that LLMs defy a binary categorization as either open- or closed-source. Due to the complexity of its architecture and training processes, LLMs should instead be positioned on a multidimensional openness chart, representing each LLM component individually with respect to its openness level (Fig. 1). The correct openness representation is essential not only for LLM developers and AI researchers but also for informing public debate, enabling all interested parties to better understand and critically assess the varying degrees of openness and the implications these have for a sensitive field like medicine.

How do open-source models perform compared to proprietary closed models in medicine?

Over the last few years, an ever-evolving ecosystem of general-purpose and medical domain-specific LLMs has emerged[15], and it is increasingly difficult to keep track of all new model releases and their performance. Closed-source LLMs, such as GPT-4o[12] and Claude 3.5 Sonnet[16], are well-known to the public and regularly used in personal and business environments. In contrast, more open models such as the OLMo 7B Instruct[14] and BigScience Large Open-science Open-access Multilingual Language Model (BLOOM)[17] are only known to AI professionals and LLM enthusiasts. Open-weight models, such as Mistral 7B[13] have recently gained popularity due to their ability to run efficiently on edge devices such as consumer laptops and mobile phones. Today, even small and open LLMs in the range of seven to eight billion parameters, which can be operated efficiently on consumer laptops, outperform OpenAI's original ChatGPT powered by GPT-3.5[11], which was perceived as a major breakthrough in 2022 and sparked massive excitement about a new wave of GenAI. In July 2024, the open-weight model Llama 3.1 405B[18] nearly closed the gap to closed-source models such as GPT-4o at least on the Measuring Massive Multitask Language Understanding (MMLU) benchmark[19] and other established LLM benchmarks[18].

Benchmarking LLMs is still notoriously difficult, in particular in medicine, where meaningful evaluation methods are still missing.

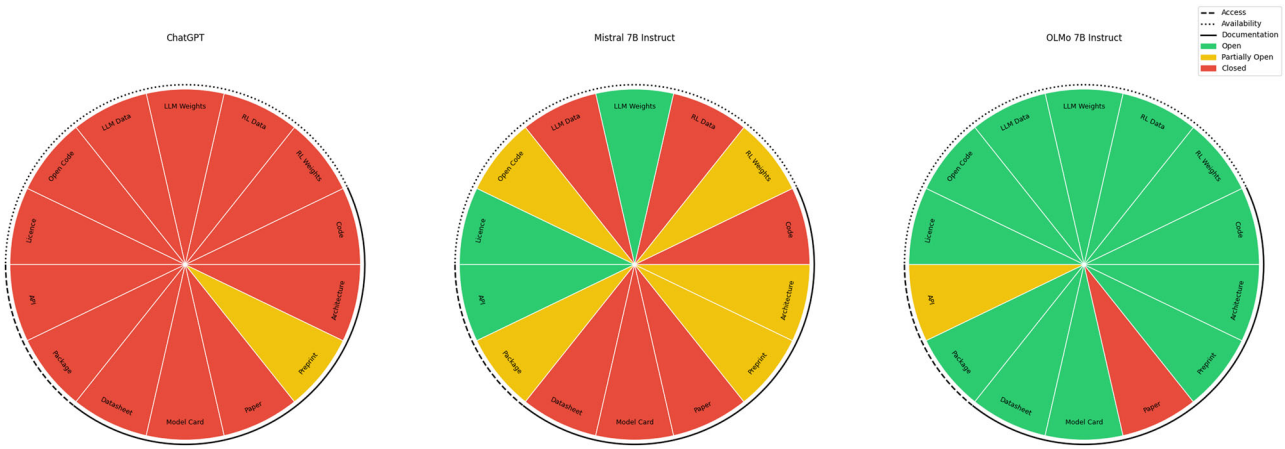

Fig. 1 | Degrees of openness with three representative LLMs: ChatGPT (GPT-3.5), Mistral 7B Instruct, and OLMo 7B Instruct. Each chart measures 14 key criteria with respect to three main categories: access, availability, and documentation. Subcategories include the openness of the code base, data, weights, and license. In terms of documentation, the code and architecture must be well-described, with a preprint, peer-reviewed paper, model card, and data sheet for additional details. Finally, access is measured by the availability of an installable Python package and an API for testing purposes. All of these criteria are measured according to three levels: open (green), partially open (yellow), and closed (red). ChatGPT shows predominantly closed access, whereas OLMo 7B Instruct demonstrates openness across most criteria. Mistral 7B Instruct balances open and restricted features, with open access to model weights but hiding details about the process they used to create it. Framework and data adopted from Liesenfeld et al.[9]. More detailed information about the framework criteria and individual LLMs data can be found here: http://opening-up-chatgpt.github.io.

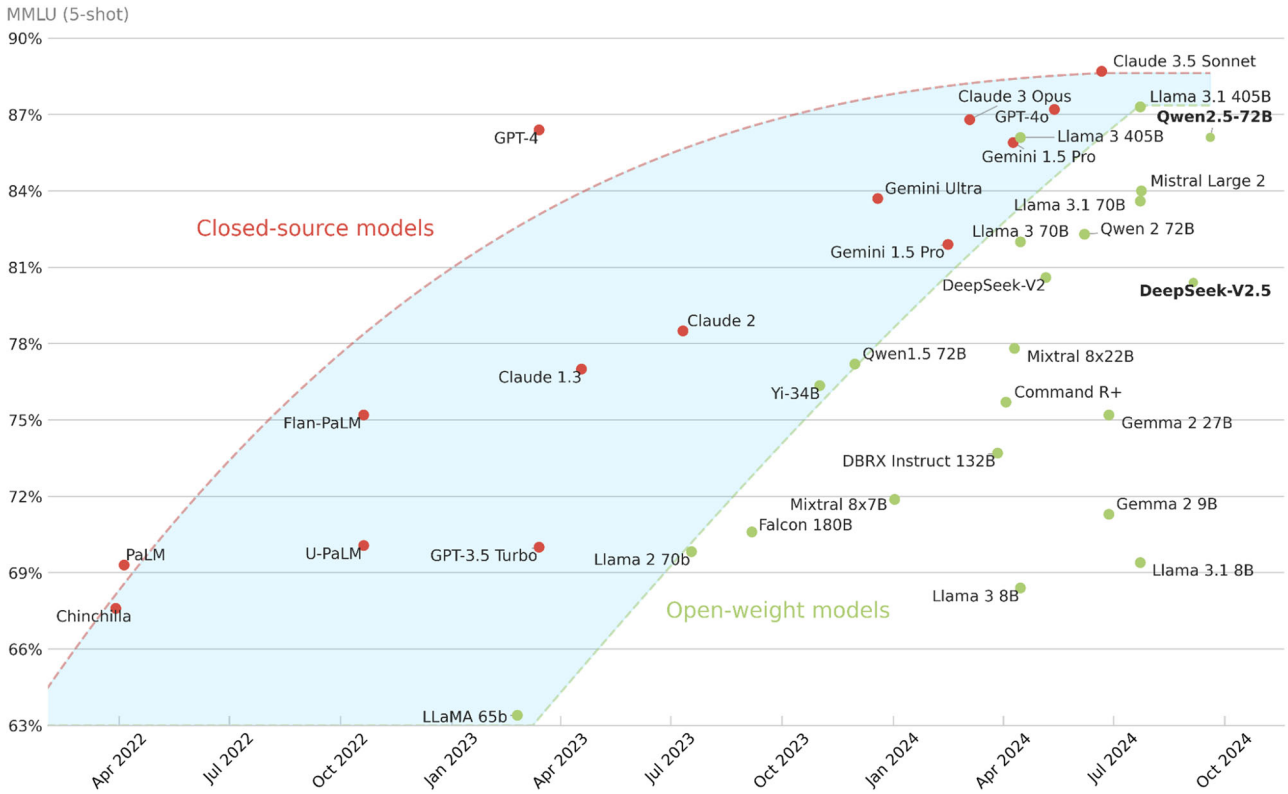

Fig. 2 | Performance comparison of closed-source and open-weight large language models on the MMLU (5-shot) benchmark. The narrowing performance gap on this popular benchmark shows how open-weight models have progressively caught up with closed-source counterparts. This evaluation consists of 15,908 questions across 57 different academic subjects (e.g., virology, business ethics, astronomy). For comparison, the authors of MMLU estimate that human domain experts achieve around 89.8% accuracy[19]. However, MMLU only measures performance in terms of general knowledge and does not target capabilities like logical reasoning, code, multilingualism, and helpfulness. To provide a better overview of general performance, it must be completed with a suite of diverse and high-quality benchmarks. Data were collected based on scores reported by model developers from April 2022 to August 2024.

Automatic, unbiased LLM evaluation methods are highly desirable but all attempts to reliably automate this time-consuming process for medical tasks have failed. Human LLM output evaluation executed by domain experts remains the gold standard. Moving forward, LLMs will increasingly be integrated into medical devices, and therefore it might be advisable to not focus the assessment on the domain-specific medical knowledge of the LLM but rather on the models' cognitive core capability and its interaction with other system components[20]. Acknowledging the current challenge in meaningful measuring LLM performance, current data imply that closed-source models perform slightly better than their open-source counterparts at least on the current medical benchmarks. However, in the recent past, open-weight models have always caught up with their closed-source counterparts within just a few months (Fig. 2). Therefore, it seems that closed-source models do not have an inherent advantage over open-source models but rather a head start, providing limited evidence to suggest that closed-source LLM approaches in medicine will be favored in the long-term. As the performance gap narrows, medical community engagement is becoming an essential advantage of open-source models. As recently shown, the collaboration of researchers and medical experts in fine-tuning open LLMs can achieve performance levels in medical text summarization comparable to proprietary models, while still providing benefits in transparency and customization[21].

Summary

The claimed systematic dangers of misuse of powerful LLMs. Those who advocate for closed-source models have largely based their arguments on non-domain-specific LLMs and on the basis of the need to avoid systemic risks that could accompany the misuse of these general-purpose models. The core argument is that the power of closed-source models is so great that they can be used for large-scale disinformation campaigns, cyber-attacks, or assist in the creation of biological weapons by terrorists[22]. Several companies leading the development and deployment of closed general-purpose LLMs have lobbied for immediate legislation and strict LLM regulation by government authorities[23]. Conversely, another group of big-tech companies and the open-source community, which are developing more open models, have been lobbying in the opposite direction, to ensure that research and development are not blocked by too restrictive legislation[24]. The European Union (EU) has taken a critical step towards regulating AI with the adoption of the AI Act on 13 March 2024. The EU AI Act presumes a GPAI model poses a systemic risk and imposes a variety of requirements for models trained with more than 10^{25} floating point operations (FLOPs)[25]. The EU Commission defines a FLOP as a proxy for model capabilities, and the exact FLOP threshold can be updated upwards or downwards by the Commission, e.g. in the light of progress in objectively measuring model capabilities and of developments in the computing power needed for a given performance level[25]. The EU legislative approach

has been based on the premise that the capabilities of models above this threshold are not yet well enough understood. They could pose systemic risks, which is why it is reasonable to subject their providers to an additional set of obligations[25]. Although training these models requires enormous computational resources (e.g., 3.8×10^{25} FLOPs for Llama 3.1 405B[18]), the process of removing content restrictions can be accomplished with far less computational expense. This can be performed through fine-tuning, as well as cheap weight modification techniques like abliteration, requiring less than 0.001% of the training budget[26]. These uncensored models can then be operated by anyone with access to sufficient GPUs. In other words, there is evidence that the current EU Act approach may simply function to restrict the operators of large LLMs for responsible purposes, while having no effect on semi-organized actors who would wish to operate models for irresponsible or illegal purposes.

How should policy proceed? On the one hand, closed LLMs provide very little of what is needed for the legal application of these approaches in downstream medical decision support systems, and, on the other hand, through its practical implications for very large models (and their labeling as having systemic risks), the EU AI makes it hard for truly open large LLMs to be made available. The US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023) may have similar effects, depending on outcomes and implementation after a public consultation. In the view of the authors, at least for now, the limitations of LLMs with respect to their reasoning and planning make a rapid and direct evolution from an LLM to an uncontrollable artificial general intelligence (AGI) highly unlikely. Additionally, an informed society is rapidly adapting to AI-generated text, images, and videos, which reduces the receptivity to misinformation via AI-generated content. Hence, it seems that rather than advocating for immediate strict regulation and bans, efficient theoretical education and practical training on transparent AI systems are urgently needed for healthcare workers and the general public. To facilitate this, specific living labs for LLMs in healthcare can serve as an ideal collaborative environment, where patients and interested parties from technology, healthcare, ethics, regulation, and policy co-create open LLMs and downstream LLM-enabled tools in real clinical settings, and collaboratively develop actionable guidelines for the deployment of these open LLMs in medicine within an existing legal framework (and even facilitating the further refinement of these frameworks).

Weighing up the relative merits of open and closed LLMs in medicine, current closed systems are inappropriate for medicine due to their lack of transparency and their exclusion of effective quality and safety control. These problems could be resolved in future models which are closed to the public and have access only on limited contractual terms for the developers of clinical decision support systems. However, if LLMs will have such a profound influence on the delivery of medicine as some claim, surely, like the libraries and their medical textbooks and journals of the past, the public and healthcare professionals should have access to the very information (and its basis) that is dictating the practice of medicine. If information is closed, it is not really information, and should not be used for informing anything, let alone healthcare. Some of the proponents of closed LLMs are labeling them as a general-purpose technology that can revolutionize human society and industry, facilitating the spread of information and a new Enlightenment with a consequent era of economic growth. The power of the printing press was the power of an open foundational technology.

Data availability

No datasets were generated or analyzed during the current study.

Lars Riedemann[1] ✉, Maxime Labonne[2] & Stephen Gilbert[3]
[1]Department of Neurology, Heidelberg University Hospital, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany. [2]Liquid AI, Inc., 314 Main St., Cambridge, MA, 02142, USA. [3]Else Kröner Fresenius Center for Digital Health, TUD Dresden University of Technology, Fetscherstr. 74, 01307 Dresden, Germany.
✉e-mail: lars.riedemann@med.uni-heidelberg.de

Received: 8 October 2024; Accepted: 14 November 2024;
Published online: 27 November 2024

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Koch, M.-C. UKE: AI writes electronic doctor's letters. *Heise online* https://www.heise.de/en/news/UKE-AI-writes-electronic-doctor-s-letters-9842226.html (2024).
3. Freyer, O., Wiest, I. C., Kather, J. N. & Gilbert, S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit. Health* **6**, e662–e672 (2024).
4. Gilbert, S., Kather, J. N. & Hogan, A. Augmented non-hallucinating large language models as medical information curators. *Npj Digit. Med.* **7**, 1–5 (2024).
5. Gilbert, S. & Kather, J. N. Guardrails for the use of generalist AI in cancer care. *Nat. Rev. Cancer* **24**, 357–358 (2024).
6. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med.* **29**, 2396–2398 (2023).
7. Mahdawi, A. What is going on with ChatGPT? *The Guardian* (2024).
8. Chen, L., Zaharia, M. & Zou, J. How is ChatGPT's behavior changing over time? *arXiv* https://doi.org/10.48550/arXiv.2307.09009 (2023).
9. Liesenfeld, A., Lopez, A. & Dingemanse, M. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In: *Proceedings of the 5th International Conference on Conversational User Interfaces* 1–6 (Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3571884.3604316. (2023).
10. White, M. et al. The model openness framework: promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. *arXiv* https://arxiv.org/abs/2403.13784 (2024).
11. OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ (2022).
12. OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/ (2024).
13. Jiang, A. Q. et al. Mistral 7B. *arXiv* https://arxiv.org/abs/2310.06825 (2023).
14. Groeneveld, D. et al. OLMo: accelerating the science of language models. *arXiv* https://arxiv.org/abs/2402.00838 (2024).
15. Zhou, H. et al. A survey of large language models in medicine: progress, application, and challenge. *arXiv* https://arxiv.org/abs/2311.05112 (2024).
16. Anthropic. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.
17. BigScience Workshop, B. et al. BLOOM: a 176B-parameter open-access multilingual language model. *arXiv* https://arxiv.org/abs/2211.05100 (2023).
18. Dubey, A. et al. The Llama 3 herd of models. *arXiv* https://arxiv.org/abs/2407.21783 (2024).
19. Hendrycks, D. et al. Measuring massive multitask language understanding. *arXiv* https://doi.org/10.48550/arXiv.2009.03300 (2021).
20. Zaharia, M. et al. The shift from models to compound AI systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/ (2024).
21. Zhang, G. et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *npj Digit. Med.* **7**, 239 (2024).
22. Criddle, C. & Murgia, M. OpenAI acknowledges new models increase risk of misuse to create bioweapons. *Financial Times* (2024).
23. Kang, C. OpenAI's Sam Altman Urges A.I. Regulation in senate hearing. *The New York Times* (2023).
24. Mark Zuckerberg Stumps for 'Open Source' A.I. *The New York Times* (2024).
25. Artificial Intelligence – Q&As. *European Commission - European Commission* https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683. (2024).
26. Labonne, M. Uncensor any LLM with abliteration. https://huggingface.co/blog/mlabonne/abliteration. (2024).

Acknowledgements

S.G. received funding through the Bundesministerium für Bildung und Forschung (BMBF) for project PATH (Personal Mastery of Health & Wellness Data), which was financed through the European Union NextGenerationEU program under grant number 16KISA100K.

Author contributions

L.R. and S.G. developed the concept of the manuscript. L.R. and S.G. wrote the first draft of the manuscript. L.R., M.L., and S.G. contributed to the writing, interpretation of the content, and editing of the manuscript, revising it critically for important intellectual content. M.L. collected and plotted all data represented in Fig. 2. L.R. and S.G. had final approval of the completed version. L.R., L.M., and

S.G. have read and approved the manuscript and take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

L.R. declares shares in Shift Medical GmbH. M.L. is a Senior Staff Machine Learning Scientist for Liquid AI, Inc. S.G. declares a nonfinancial interest as an Advisory Group member of the EY-coordinated "Study on Regulatory Governance and Innovation in the field of Medical Devices" conducted on behalf of the Directorate-General for Health and Food Safety (SANTE) of the European Commission. S.G. declares the following competing financial interests: he has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd., Flo Ltd, ICURA ApS, Rock Health Inc., Thymia Ltd., FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, Directorate-General for Research and Innovation Of the European Commission, and Ada Health GmbH and holds share options in Ada Health GmbH. S.G. is a News and Views Editor for npj Digital Medicine. S.G. played no role in the internal review or decision to publish this News and Views article.

Additional information

Correspondence and requests for materials should be addressed to Lars Riedemann.