nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-022-00559-4

Developing robust benchmarks for driving forward AI innovation in healthcare

Received: 1 June 2022

Accepted: 7 October 2022

Published online: 15 November 2022

Check for updates

Diana Mincu & Subhrajit Roy

Machine learning technologies have seen increased application to the healthcare domain. The main drivers are openly available healthcare datasets, and a general interest from the community to use its powers for knowledge discovery and technological advancements in this more conservative field. However, with this additional volume comes a range of questions and concerns – are the obtained results meaningful and conclusions accurate; how do we know we have improved state of the art; is the clinical problem well defined and does the model address it? We reflect on key aspects in the end-to-end pipeline that we believe suffer the most in this space, and suggest some good practices to avoid reproducing these issues.

Our intended audience is anyone who performs benchmarking experiments on machine learning (ML) in healthcare, and submits these results to conferences or journals; and anyone reviewing for these venues. By clinical benchmarking we refer to following the combined process.

1. Choosing a problem in the healthcare space.
2. Using or creating accompanying datasets.
3. Developing a suite of ML models and their corresponding infrastructure.
4. Evaluating these models on a set of criteria for how well they solve the original problem.

The problem of finding a good benchmark is much more prevalent in the healthcare domain because there is not enough alignment on what this actually constitutes[1]. In a typical research cycle, when ML is first applied to a new medical area it leads to a publication that measures the model performance and sets the bar for that problem. The dataset, ground truth, metrics or code are not always scrutinized as much as in traditional ML, as there is a lot of interest in seeing what the new technology can actually do in practice. We would argue that benchmarking papers should be scrutinized even more, as variability in definitions, set-ups and evaluation can lead to a lot of misrepresentation of findings, as well as confusion for newcomers to the field on how to compare their work.

In the next four sections, we cover the areas where we believe most of the discrepancies between two different benchmarking publications lie: datasets, tools and practices, problem formulation and results. In our view, any such changes could have an impact on the overall clinical application, as advancements would be much easier to quantify. Inspired by ref. [2], we classify our suggestions into three categories: necessary, recommended and encouraged (Box 1–4). Each section builds upon the previous one, and is placed within a given bucket by taking into account a mix of its impact and difficulty of implementation. Even if some might not agree fully to the categorization, we expect to start a conversation around these topics that will see changes in the field.

There is existing work in this space looking at either defining reporting standards (for example STARD-AI[3], TRIPOD-AI[4]) or specifying best practices when it comes to model development and technique reporting[5]. Both of these directions encourage the inclusion of additional details in the final works to reduce uncertainty when it comes to the methods employed. We believe our work builds upon and extends these works, by looking at the end-to-end pipeline and tackling less explored topics such as tools and infrastructure.

Datasets

Large, high-quality, diverse and well documented health datasets are hard to obtain as data sharing is not the norm in ML for healthcare research[6]. First, health datasets contain extremely sensitive information and are therefore firmly regulated, with recent research looking to understand patients' attitude towards health data sharing[7,8]. This is why these datasets are typically de-identified before public release, a process that involves the removal of the patient's name, identification number, date and location of data collection. Even so, it has been recently shown that anonymized magnetic resonance imaging (MRI) or computed tomography data can be used to reconstruct a patient's face[9], which raises questions on whether the current de-identification standards are sufficient for safe public data release. Second, the collection,

Google Research, London, UK. e-mail: dmincu@google.com; subhrajitroy@google.com

BOX 1

Dataset suggestions

Necessary

- Provide a thorough description of the provenance, demographics and content of the dataset (for example, Table 1 data).
- Apply and include numerical (for example, mean, variance, min, max and correlation matrices) and/or graphical (for example, scatterplot, histogram, heatmap and dimensionality reduction) exploratory data analysis in the final work.
- Include details of how the quality of the dataset was verified by describing missing features, imbalanced data, duplicate instances, sampling bias and other dataset-specific issues.

Recommended

- Release a transparency artefact by using standardized questionnaire templates (for example, Healthsheet[20]) along with the paper.

Encouraged (private datasets only)

- Use robust infrastructure developed by non-profits such as Openmined[21] to host and manage health datasets.

BOX 2

Tools and infrastructure suggestions

Necessary

- Add an implementation section in either the main paper or the appendix.
- Add a 'How was this implementation verified?' section for submissions.

Recommended

- Add an 'Experimental environment' section in the final works, which should not count towards the page limit.

Encouraged

- Provide links to the open-source code and ways to run it.

maintenance and curation of such datasets require substantial effort, time and expense. Moreover, the datasets are seen as a competitive commercial advantage, with cases where companies are formed around the exclusive use of one. Hence, they tend to have substantial business value, thereby making it less appealing for data collectors to freely share their work. However, to broaden the impact of publishing research on non-public datasets and encourage reproducibility, data curators could set up infrastructure allowing the community to develop models by using privacy-preserving ML techniques such as federated learning[10–12]. In this setting, the data controller defines its own governance processes, associated privacy policies and access management strategies, during both the training and the validation phase. This unlocks the exploration of health datasets by external researchers while retaining data privacy, thereby accelerating progress. We do however acknowledge that there is a steep learning curve in setting this up, and it is difficult to trust such systems given that these methods are relatively new. It might be preferable then, at the beginning, for data curators to work with vetted external data scientists. Another field that is promising for retaining business value is tracing when a member of a dataset is used to train a model. Examples of this include recent work on 'watermarking' a given dataset to facilitate identification of models trained on it[13], or techniques such as membership inference[14]. We do caution that this field is not yet established, so care needs to be taken if going this route.

The low availability of public health datasets often forces the community to rally around one or two that are accessible, thereby overfocusing on a few applications. For example, popularly used electronic health records datasets such as the MIMIC-X series[15–17] primarily contain intensive care unit data, which are frequently recorded but represent merely a fraction of the population of patients admitted to hospitals. In addition, MIMIC-X is curated from a single site and is therefore less likely to produce fair and inclusive ML models. As such, models developed on these datasets may not necessarily generalize to other wards or find hospital-wide applications. A thorough characterization of the data is needed for external researchers and developers to evaluate the data quality, and decide whether the methodologies proposed in the paper can be expanded to their use case. This can also help identify unwanted demographic biases (for example skewed distribution for age, gender, socioeconomic status). This data characterization can be achieved through a breakdown of the various features inside the dataset (informally referred to as the Table 1 data[18]); we offer more suggestions on how to achieve this in the 'Necessary' section of Box 1.

Another example is that in 2020–2021 various papers trained models on datasets containing chest X-rays for COVID-19 modelling, where images of paediatric patients represented the control group. However, such models are likely to overperform as they are merely detecting children versus adults, and are inherently biased[19]. The mere availability of this dataset during the early days of the COVID-19 pandemic caused the community to overindex without careful consideration of whether it was appropriate for their research question or not. Such cases could be avoided if datasets were accompanied by a transparency artefact (for example Healthsheet[20]) along with the paper. The Healthsheet questionnaire, for example, contains specific questions on different aspects of a dataset such as the collection process, composition, distribution, cleaning and recommended use cases. Such artefacts vastly improve the transparency and accountability of datasets and data curators, and bring to the surface technical imbalances (for example improper acquisition protocol, equipment manufacturer), which may adversely affect model predictions. We realize that the creation of an extra artefact along with their manuscript is time consuming, given that researchers often work towards tight deadlines. To manage workload, researchers may choose to share such artefacts after submission.

Tools and practices

One of the more unseen and less talked about sides of ML research is the infrastructure. We use the term infrastructure to refer to the design of a system, together with the underlying code that forms it, including the environment in which it runs, and the software libraries that are used. While often it is considered unglamorous, it is absolutely necessary and can make or break the result quality and reproducibility.

While the role of good coding and system design practices has been previously explored for production or deployment set-ups[21], in-depth descriptions of the libraries or pipelines used for a publication are still lacking. This is especially important in the healthcare domain, as any impactful finding is quickly picked up by news outlets and spread as ground truth, or used by other researchers as the basis for future explorations. Looking through the retraction database for recent ML papers in healthcare that contained errors, we can find troubling examples such as "an outstanding Parkinson's Disease predictor" which the authors claim "has the potential to revolutionize the diagnosis of PD and its management". Upon later inspection it was found to exhibit errors in data, errors in methods, errors in results and errors in analyses.

BOX 3

Problem formulation suggestions

Expert-defined labels

Necessary

- Add a detailed description of the labelling process used in the paper.

Expert-guided labels

Necessary

- Add a 'Label analysis' section in the main paper.
- Investigate 'label leakage' in the data and include findings in the appendix or supplementary information.

Recommended

- Implement a multistage label quality framework consisting of manual feature inspection, label statistics and case reviews.

BOX 4

Results suggestions

Necessary

- Include fairness measurements, calibration scores and label-dependent metrics during model evaluation.
- Include comparisons with baseline models and tune the bias–variance trade-off with respect to model complexity.

Recommended

- Perform failure analysis — identify instances where the model fails and investigate their commonalities. We recommend methods such as the 'medical algorithmic audit' framework for structured failure analysis[45].

Encouraged

- Include thorough descriptions of experiments that need to be done, but were not performed.
- Add model visualizations to the resulting research.

All of these could be caught earlier in the review process if a detailed 'infrastructure used' section raised any red flags. We would like to encourage authors to describe their implementation and system design in more depth. Adding a detailed diagram of their modelling pipeline or including rigorous descriptions of data processing modules, such as mapping tools that might have been developed, are a few examples of what we would expect to see. This is a lightweight proposal and should be fairly simple to adopt, with only a small overhead for the authors.

To enforce this more broadly, we would like to encourage conferences to add a 'How was this implementation verified?' section for submissions. Authors will be required to describe what steps they have taken to ensure the correctness of their work. Where appropriate, code reviews are a great way for any work to be sanity checked, and while not foolproof this can still help catch problems before it is too late. Furthermore, we want to advocate for adding unit tests and checking for code coverage. Apart from the immediate benefit of validating the expected behaviour, it is much quicker to understand a piece of code by looking at how it is used in practice, and tests offer a glimpse into this. While our recommendations rely heavily on coding, we acknowledge that certain techniques reuse existing implementations. Alternative ways of testing, such as checking for a matching performance to the original work, would also be covered. While the addition of this section introduces an overhead for both the reviewers and the authors, it contributes to the overall reproducibility goal and propagates good practices throughout the community.

Even so, it takes time and resources to investigate the accuracy of reported findings once they have been publicly released, and often this involves rebuilding the paper's entire set-up from scratch. To make this easier, conferences and journals have started to include a section on code availability, encouraging researchers to open-source their work. This is a great step forward, and even though still not widely adopted the importance of code publication is accepted by the research community. As an example, the Machine Learning for Health (ML4H) conference collects statistics on how many submissions will have their code released. In the year 2020, only 66% of submissions reported they would do so. This number increased in 2021 to 73%.

We believe that, in addition to this, conferences should ask for an extra section called 'Experimental environment', which should not count towards the page limit. This would be a superset of the 'Code availability' tick box, requiring authors to also list all publicly available libraries used and their version. We believe this is important as familiarity with the tools used is a big factor for trustworthiness. We understand that this could be a cumbersome task at first, especially as projects grow larger and there could be tens if not hundreds of libraries used, but without it there is no true reproducibility.

Above all, open-sourcing the code remains the most transparent way for the community to check results. This is strengthened if it is released together with a script to run the code, and real or synthetic data depending on the possibilities. In the case of synthetic data we would also refer back to the dataset section for further recommendations.

Problem formulation

We focus on clinical problems that have been posed as supervised prediction problems as these constitute the majority of the ML for healthcare literature.

The most important step for supervised learning in healthcare is to decide which clinical labels to predict. Error or bias in labelling is common in ML and can lead to subpar models. It was estimated that test sets of popular datasets contain at least 3.3% label errors on average[22]. Correcting these labels allows lower-capacity models to outperform commonly reported state-of-the-art models.

Proper annotation of instances in healthcare datasets usually depends on the expert knowledge of medical professionals. Labels are typically either fully defined by clinicians, or generated semi-autonomously using rule-based methods incorporating clinical guidance. Examples of the former include skin classification from dermatology imaging[23], breast lesion detection in mammograms[24], referral recommendation in optical coherence tomography[25], segmentation of lymph nodes on multiparametric MRI[26] and seizure detection using EEG data[27]. Expert-guided labels usually involve developing a set of rules to identify certain conditions and using the rule set to annotate the full dataset. Examples include prediction of adverse events or interventions in electronic health record data such as acute kidney injury[28], mechanical ventilation[29], medication orders[30] and continuous renal replacement therapy[31]. In both cases, the labels would closely mirror a clinician's workflow, the goal of labelling being to document the process in which medical professionals make decisions. We therefore strongly encourage the use or improvement of existing labels such as those in the Phenotype KnowledgeBase[32].

For expert-defined labels it is imperative that a detailed description of the labelling process used is included in the paper. Note that even when labels are fully defined by experts there may be variability among healthcare professionals on the annotation of an instance.

Researchers should report whether a single clinician/expert or a committee of experts labelled instances, and if the latter then report the inter-rater agreement. To demonstrate thoroughness, authors should report the average time it took to annotate each instance. Researchers may also provide a benchmark by sharing the human-level performance. Note that these suggestions are not exhaustive and are included to guide the researchers. This reporting will capture the subjectivity in labelling among raters, and provide an idea of the robustness and reliability of the labelling process. It also sets a bar for how subsequent studies should approach labelling for other tasks defined on this dataset.

In the case of rule-based or expert-guided labels, a robust process is required to validate them, as they often contain anomalies in individual instances and/or suffer from label leakage. We suggest performing an analysis on the distribution for each label, including patient demographics for cohorts corresponding to each label class, label counts per subject or instance, and distribution statistics (mean, median, percentiles, variance). In addition, for continuous labels in temporal data, the distribution of the label onset time and distribution of the label duration should also be reported. These should be cross-referenced with expert clinicians to detect any anomalies in the label distribution.

We also suggest that researchers investigate whether there is any potential label leakage in their problem formulation. This usually leads to a false high performance and requires domain knowledge to identify and solve. Label leakage may happen for various reasons when data from either the validation or the test set have leaked into the training set. This problem can be relatively easily solved by checking whether the same instances exist in multiple splits and if there is a duplication of instances, and by ensuring that the blind test set remains locked until the final results are computed for inclusion in the paper. Label leakage may also occur when certain operational or observational features undesirably reveal the state of a label. Researchers should perform feature importance analysis to inspect suspicious relationships[33]. If identified, such features should be reviewed with clinicians to identify whether they are indeed undesirably indicative of the ground truth.

Ideally this would be combined in a multistage label quality framework consisting of manual feature inspection, label statistics and case reviews. A methodical approach allows researchers to ensure consistency throughout the process. This approach, albeit time consuming, if open-sourced can be adapted by the community on other tasks on the same dataset, or even datasets from other domains, reducing the workload in the long run.

Results

Investigating and comparing model results becomes a make-or-break step, as the ultimate goal for much healthcare research is aiding clinical practice in some capacity. For this to take place, we need confidence that the model will not cause any harm – either by making the current state worse, or by introducing any new problems. Moreover, the further away from clinical practice a proposed method is, the more evidence we need that it actually works.

A growing field has been looking at fairness and robustness evaluation for ML in healthcare, and a number of works have been advocating for more fairness metrics to be included with model reporting. One way to do this would be by making use of model evaluation tools such as TensorFlow Model Analysis. In addition to adding a layer of consistency when it comes to analysis, such tools have the added benefit of providing APIs (application programming interfaces) for fairness measurements. By using and reporting these results, it can become common practice to look beyond the full test-set performance. Comprehensively reporting a broad set of metrics will allow different aspects of the model to be questioned and understood. For example, looking at the class imbalance and showcasing metrics based on the label skew is critical (for example reporting the area under the precision–recall curve alongside just the area under the receiver operating characteristic), as well as including clinically relevant metrics such as sensitivity and specificity[34].

When it comes to fairness and robustness, there are a few key issues that keep surfacing: (1) performance across subgroups differs; (2) similarly performing models behave differently in unexpected ways when there is a shift from the training distribution. Recent work has shown that general mitigation techniques developed for some fairness issues do not translate so well when it comes to healthcare applications[35]. Together with ref. [36], it showcases a number of stress tests that were performed during model investigation, that we argue should be performed as part of the usual benchmarking routine to bring such issues to the surface before it is too late. A popular benchmarking study on MIMIC-III[1] has recently been found to display issues when it comes to fairness and generalizability[37]. We therefore want to stress the importance of the community becoming more familiar with the model's performance in different contexts, and include stress tests. Future improvements could then not only target the base model performance on the training set, but also see what technique is the most resilient when faced with real-world contexts.

Aside from looking at metrics and tables, visualizations can also help investigate the model's performance. A few suggestions would be activation atlases[38], attention heatmaps[39], grand tour[40], integrated gradients[41] or concept activation vectors[42]. These can help identify what the model is learning and help test these techniques in different contexts, providing valuable data for future research directions. We do want to acknowledge that, in the fields of model explainability and interpretability, results can be misinterpreted[43,44], and urge researchers to familiarize themselves with the various techniques and their failure modes to avoid misuse.

Finally, we are aware that there is always more work left to be done when finalizing a research paper. More often than not, there are lingering experiments that authors wanted to perform, but were unable to owing to various constraints. While some are listed in the limitations sections, typically these address continuations of experiments already mentioned. We believe that asking authors to further write down precise experiments that were left out can help both expand upon that work, and also spread awareness of key tests.

Clinical deployment

While innovative ML models have been developed for healthcare, very few of them find real-world application[46,45]. Recent surveys on ML-based clinical tools have shown that well validated models, achieving good performance in the development stage, may fail to show any clinical benefit for patients when compared with routine care[47].

We acknowledge that the deployment of ML in healthcare for researchers is hard, as barriers to implementation include regulation, incentives, lack of appreciation and generalizability concerns, to name a few. In addition, prospective validation studies require time and money, which can be a big challenge. Under these circumstances, papers that go the extra mile[47,48] and show some form of validation studies should be positively distinguished.

Papers studying the clinical effectiveness of ML tools should be rigorous in reporting various aspects of the study, including but not limited to study setting, criteria for inclusion, human–algorithm interaction and its downstream effects, methods for continual learning and most importantly a comparison with existing clinical practice. To improve the quality of reporting, we recommend that authors follow validated guidelines such as checklists published by the CONSORT-AI and SPIRIT-AI steering groups[45,49].

While current benchmarking papers are more focused on creating an upstream data science benchmark for clinical research in healthcare, we strongly believe that the future of applied healthcare research will see a lot more emphasis on the clinical deployment aspect, as the field moves from theory to practice and the array of challenges associated with it are explored in greater depth[50,51].

References

1. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 96 (2019).
2. Heil, B. et al. Reproducibility standards for machine learning in the life sciences. *Nat. Methods* **18**, 1132–1135 (2021).
3. Viknesh, S. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat. Med.* **26**, 807–808 (2020).
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. Kakarmath, S. et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. *npj Digit. Med.* **3**, 134 (2020).
6. Hulsen, T. Sharing is caring—data sharing initiatives in healthcare. *Int. J. Environ. Res. Public Health* **17**, 3046 (2020).
7. Atkin, C. et al. Perceptions of anonymised data use and awareness of the NHS data opt-out amongst patients, carers and healthcare staff. *Res. Involv. Engagem.* **7**, 40 (2021).
8. Chico, V., Hunn, A. & Taylor, M. *Public Views on Sharing Anonymised Patient-Level Data Where There Is a Mixed Public and Private Benefit* (Univ. Melbourne, 2019).
9. Schwarz, C. G. et al. Identification of anonymous MRI research participants with face-recognition software. *New Engl. J. Med.* **381**, 1684–1686 (2019).
10. Rieke, N. et al. The future of digital health with federated learning. *npj Digit. Med.* **3**, 119 (2020).
11. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
12. Ngong, I. Maintaining privacy in medical data with differential privacy. *OpenMined Blog* https://blog.openmined.org/maintaining-privacy-in-medical-data-with-differential-privacy/ (2020).
13. Sablayrolles, A., Douze, M., Schmid, C. & Jegou, H. Radioactive data: tracing through training. *Proc. Mach. Learning Res.* **119**, 8326–8335 (2020).
14. Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y. & Jegou, H. White-box vs black-box: Bayes optimal strategies for membership inference. *Proc. Mach. Learning Res.* **97**, 5558–5567 (2019).
15. Johnson, A. et al. MIMIC-IV (version 1.0) *PhysioNet* https://doi.org/10.13026/s6n6-xd98 (2021).
16. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
17. Lee, J. et al. Open-access MIMIC-II database for intensive care research. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2011**, 8315–8318 (2011).
18. Hayes-Larson, E., Kezios, K., Mooney, S. & Lovasi, G. Who is in this study, anyway? Guidelines for a useful Table 1. *J. Clin. Epidemiol.* **114**, 125–132 (2019).
19. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
20. Rostamzadeh, N. et al. Healthsheet: development of a transparency artifact for health datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency* 1943–1961 (Association for Computing Machinery, 2022).
21. Sculley, D. et al. Hidden technical debt in machine learning systems. *Adv. Neural Inf. Process. Syst.* **28**, 2503–2511 (2015).
22. Northcutt, C., Athalye, A. & Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks 1* (2021).
23. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
24. Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).
25. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
26. Zhao, X. et al. Deep learning-based fully automated detection and segmentation of lymph nodes on multiparametric-MRI for rectal cancer: a multicentre study. *eBioMedicine* **56**, 102780 (2020).
27. Roy, S. et al. Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data. *eBioMedicine* **66**, 103275 (2021).
28. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
29. Wang, S. et al. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proc. ACM Conference on Health, Inference, and Learning* 222–235 (Association for Computing Machinery, 2020).
30. Rough, K. et al. Predicting inpatient medication orders from electronic health record data. *Clin. Pharmacol. Ther.* **108**, 145–154 (2020).
31. Roy, S. et al. Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. *J. Am. Med. Inform. Assoc.* **28**, 1936–1946 (2021).
32. Kirby, J. C. et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* **23**, 1046–1052 (2016).
33. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **6**, 15 (2012).
34. Hicks, S. A. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **12**, 12 (2022).
35. Schrouff, J. et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? Preprint at *arXiv* https://arxiv.org/abs/2202.01034 (2022).
36. D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* **23**, 1–61 (2022).
37. Röösli, E., Bozkurt, S. & Hernandez-Boussard, T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Sci. Data* **9**, 24 (2022).
38. Carter, S., Armstrong, Z., Schubert, L., Johnson, I. & Olah, C. Exploring neural networks with activation atlases. *Distill* https://distill.pub/2019/activation-atlas/ (2019).
39. Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T. & Blunsom, P. Reasoning about entailment with neural attention. Preprint at *arXiv* https://arxiv.org/abs/1509.06664 (2016).
40. Li, M., Zhao, Z. & Scheidegger, C. Visualizing neural networks with the grand tour. *Distill* https://distill.pub/2020/grand-tour/ (2020).
41. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *Proceedings of the 34 th International Conference on Machine Learning, PMLR* https://doi.org/10.48550/arXiv.1703.01365 (2017).
42. Mincu, D. et al. Concept-based model explanations for electronic health records. In *Proc. Conference on Health, Inference, and Learning* 36–46 (Association for Computing Machinery, 2021).
43. Adebayo, J. et al. Sanity checks for saliency maps. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems* (2018).
44. Arun, N. et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.* **3**, e200267 (2021).

45. Liu, X. et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.* **25**, 1467–1468 (2019).
46. Lu, C. et al. Deploying clinical machine learning? Consider the following…. Preprint at *arXiv* https://arxiv.org/abs/2109.06919 (2021).
47. Zhou, Q., Chen, Z. H., Cao, Y. H. & Peng, S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *npj Digit. Med.* **4**, 12 (2021).
48. Biswal, S. et al. SLEEPNET: automated sleep staging system via deep learning. Preprint at *arXiv* https://arxiv.org/abs/1707.08262 (2017).
49. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
50. Ryffel, T. et al. A generic framework for privacy preserving deep learning. Preprint at *arXiv* https://arxiv.org/abs/1811.04017 (2018).
51. Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K. & Oakden-Rayner, L. The medical algorithmic audit. *Lancet Digit. Health* **4**, e384–e397 (2022).

Acknowledgements

We thank the clinicians who offered their help and opinions when reviewing this paper: L. Hartsell and M. Seneviratne. We also thank our colleagues and collaborators, N. Tomasev, K. Heller, J. Schrouff, N. Rostamzadeh, C. Ghate, L. Proleev, L. Hartsel, N. Broestl, G. Flores and S. Pfohl, for their help and support in reviewing and beta-testing our opinions.

Competing interests

Both authors are employed by Google UK.

Additional information

Correspondence should be addressed to Diana Mincu or Subhrajit Roy.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.