**comment**

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals

A comprehensive overview of medical AI devices approved by the US Food and Drug Administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho and James Zou

Medical artificial-intelligence (AI) algorithms are being increasingly proposed for the assessment and care of patients. Although the academic community has started to develop reporting guidelines for AI clinical trials[1–3], there are no established best practices for evaluating commercially available algorithms to ensure their reliability and safety. The path to safe and robust clinical AI requires that important regulatory questions be addressed. Are medical devices able to demonstrate performance that can be generalized to the entire intended population? Are commonly faced shortcomings of AI (overfitting to training data, vulnerability to data shifts, and bias against underrepresented patient subgroups) adequately quantified and addressed?

In the USA, the US Food and Drug Administration (FDA) is responsible for approving commercially marketed medical AI devices. The FDA releases publicly available information on approved devices in the form of a summary document that generally contains information about the device description, indications for use, and performance data of the device's evaluation study. The FDA has recently called for improvement of test-data quality, improvement of trust and transparency with users, monitoring of algorithmic performance and bias on the intended population, and testing with clinicians in the loop[4,5]. To understand the extent to which these concerns are addressed in practice, we have created an annotated database of FDA-approved medical AI devices and systematically analyzed how these devices were evaluated before approval. Additionally, we have conducted a case study of pneumothorax-triage devices and found that evaluating deep-learning models at a single site alone, which is often done, can mask weaknesses in the models and lead to worse performance across sites.
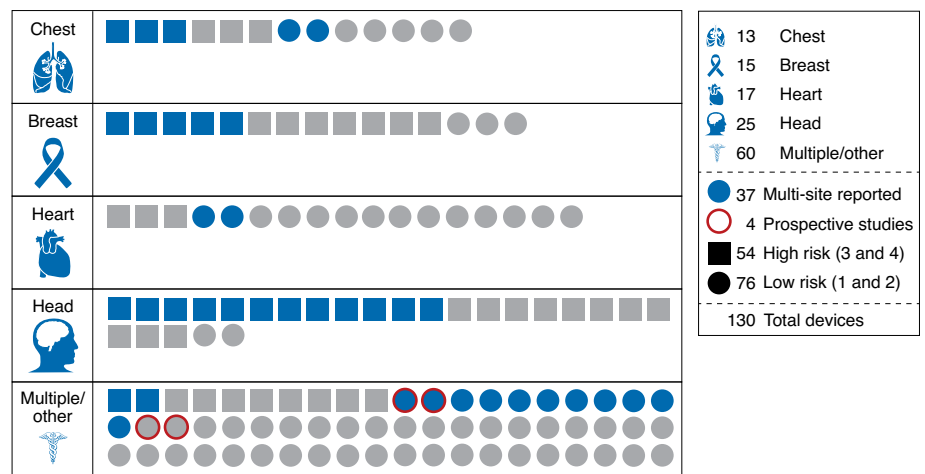
Fig. 1 | Breakdown of 130 FDA-approved medical AI devices by body area. Devices are categorized by risk level (square, high risk; circle, low risk). Blue indicates that a multi-site evaluation was reported; otherwise, symbols are gray. Red outline indicates a prospective study (key, right margin). Numbers in key indicate the number of devices with each characteristic.

Curating a comprehensive database of FDA-approved medical AI

We aggregated all of the medical AI devices approved by the FDA between January 2015 and December 2020 (refs. [6–8]). Because searching for specific terms is not possible on the FDA website (https://www.fda.gov/)[9], we downloaded the PDF file for the summary document of each approved device, extracted the text, and searched for AI keywords to create our initial corpus. We then merged this corpus with two existing databases of FDA-approved AI devices[9,10] and filtered for AI relevance to create a comprehensive database (https://ericwu09.github.io/medical-ai-evaluation).

From the summary document of each device, we extracted the following information about how the algorithm was evaluated: the number of patients enrolled in the evaluation study; the number of sites used in the evaluation; whether the test data

were collected and evaluated concurrently with device deployment (prospective) or the test set was collected before device deployment (retrospective); and whether stratified performance by disease subtypes or across demographic subgroups was reported. Additionally, we assigned a risk level from 1 to 4 to each device (1 and 2 indicate low risk; 3 and 4 indicate high risk) according to guidelines from an FDA proposal[4]. In total, we compiled 130 approved devices that met our review criteria. We present a compilation of all the devices, organized by body area, risk level, prospective/retrospective studies, and multi-site evaluation (Fig. 1).

Most evaluations perform only retrospective studies

Almost all of the AI devices (126 of 130) underwent only retrospective studies at their submission, based on the FDA summaries.

None of the 54 high-risk devices were evaluated by prospective studies. For most devices, the test data for the retrospective studies were collected from clinical sites before evaluation, and the endpoints measured did not involve a side-by-side comparison of clinicians' performances with and without AI.

More prospective studies are needed for full characterization of the impact of the AI decision tool on clinical practice, which is important, because human–computer interaction can deviate substantially from a model's intended use. For example, most computer-aided detection diagnostic devices are intended to be decision-support tools rather than primary diagnostic tools. A prospective randomized study may reveal that clinicians are misusing this tool for primary diagnosis and that outcomes are different from what would be expected if the tool were used for decision support.

The number of evaluation sites and samples are often not reported

Among the 130 devices we analyzed, 93 devices did not have publicly reported multi-site assessment included as a part of the evaluation study. Of the 41 devices with the number of evaluation sites reported, 4 devices were evaluated in only one site, and 8 devices were evaluated in only two sites. This suggests that a substantial proportion of approved devices might have been evaluated only at a small number of sites, which often tend to have limited geographic diversity[11]. The number of approvals for AI devices has increased rapidly in the past 5 years, with over 75% of approvals coming in the past 2 years and over 50% coming in the past year. However, the proportion of approvals with multi-site evaluation and reported sample size has remained stagnant during the same period of time. Furthermore, the published reports for 59 devices (45%) did not include the sample size of the studies. Of the 71 device studies that had this information, the median evaluation sample size was 300. Only 17 device studies reported that demographic subgroup performance was considered in their evaluations.

Although the number of sites used in a study is available to the FDA, it is also important that this information be consistently reported in the public summary document in order for clinicians, researchers, and patients to make informed judgments about the reliability of the algorithm. Multi-site evaluations are important for the understanding of algorithmic bias and reliability, and can help in accounting for variations in the equipment used, technician standards, image-storage formats, demographic makeup, and disease prevalence.

Case study of multi-site evaluation for pneumothorax detection

As it is critical to understand how a model's performance can be generalized to a broad and diverse population, we explored how AI models might perform when evaluated on patients from multiple clinical sites that represent different populations. To this end, we chose the detection of pneumothorax (collapsed lung) as a case study, as there are currently four medical devices cleared via section 510(k) of the Food, Drug and Cosmetic Act that are approved for the triage of X-ray images for the presence of pneumothorax, and there are multiple publicly available chest X-ray datasets that include pneumothorax as a condition. We used three datasets, each from a different hospital site in the USA: the National Institutes of Health Clinical Center in Bethesda, Maryland (NIH)[12]; Stanford Health Care in Palo Alto, California (SHC)[13]; and Beth Israel Deaconess Medical Center in Boston, Massachusetts (BIDMC)[14]. We used a DenseNet-121 deep-learning architecture[15] that has been demonstrated to be a top-performing model for the classification of chest conditions[13,16].

To quantify how the AI's performance varies across sites, we trained three separate deep-learning models on data from patients at each of the three sites and then evaluated the models on the test set from the other two sites. Each model takes as input a chest X-ray image and makes a binary prediction for pneumothorax. A summary of the results shows, for example, the performance of the model trained on SHC when evaluated on the test patients at SHC (distinct from the SHC training set), BIDMC, and NIH (Table 1, top row). We found that the algorithm trained on the NIH data achieved good performance on independent NIH test patients (area under the receiver operating characteristic curve (AUC), 0.883) but performed much worse on the BIDMC test patients (AUC, 0.759) and SCH test patients (AUC, 0.779) (Table 1). Across the board, we found substantial drop-offs in model performance when the models were evaluated on a different site. While the within-site test AUC remained high (average, 0.893), the performance degraded markedly by an average of 0.072 AUC and up to 0.124 AUC when evaluated on the other two sites (Table 1). Some of the performance variations could have been due to differences in patient demographics across sites. For example, when evaluating each of the three models on the BIDMC test set, we found that the performance disparity between white and Black patients increased from 0.024 AUC with the BIDMC-trained model to 0.043 AUC and 0.109 AUC with the other two models.

Recommendations

Evaluating the performance of AI devices in multiple clinical sites is important for ensuring that the algorithms perform well across representative populations. Encouraging prospective studies with comparison to standard of care reduces the risk of harmful overfitting and more accurately captures true clinical outcomes. Post-market surveillance of AI devices is also needed for understanding and measurement of unintended outcomes and biases that are not detected in prospective, multi-center trials[17,18]. ❒

Table 1 | Cross-site performance of an algorithm

Site	SHC (N = 18688)	BIDMC (N = 23204)	NIH (N = 11196)
SHC	**0.903 ± 0.009**	0.870 ± 0.012	0.852 ± 0.020
BIDMC	0.827 ± 0.012	**0.892 ± 0.009**	0.839 ± 0.021
NIH	0.779 ± 0.013	0.759 ± 0.016	**0.883 ± 0.015**

Each row represents an algorithm trained on a single site; columns indicate the dataset the algorithm was evaluated on. Each cell contains the AUC and 95% confidence interval. Bolded numbers indicate within-site performance. The data size (N) refers to the test dataset.

Eric Wu[1], Kevin Wu[2], Roxana Daneshjou [2,3], David Ouyang[4], Daniel E. Ho[5] and James Zou [1,2,6]✉

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [3]Department of Dermatology, Stanford University, Stanford, CA, USA. [4]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [5]Stanford Law School, Stanford University, Stanford, CA, USA. [6]Chan-Zuckerberg Biohub, San Francisco, CA, USA. ✉e-mail: jamesz@stanford.edu

References

1. Liu, X. et al. *Nat. Med.* **26**, 1364–1374 (2020).
2. Norgeot, B. et al. *Nat. Med.* **26**, 1320–1324 (2020).
3. Cruz Rivera, S., Liu, X. & Chan, A. W. et al. *Nat. Med.* **26**,

1351–1363 (2020).
4. U.S. Food & Drug Administration Center for Devices and Radiological Health. https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm514737.pdf (October 2017).
5. U.S. Food & Drug Administration Center for Devices and Radiological Health. https://www.fda.gov/media/145022/download (January 2021).
6. U.S. Food & Drug Administration. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm (accessed 14 December 2020).
7. U.S. Food & Drug Administration. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm (accessed 14 December 2020).
8. U.S. Food & Drug Administration. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm (accessed 14 December 2020).
9. Benjamens, S., Dhunnoo, P. & Meskó, B. Npj Digit. Med. 3, 1–8 (2020).
10. American College of Radiology. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms (accessed 29 November 2020).

11. Kaushal, A., Altman, R. & Langlotz, C. J. Am. Med. Assoc. 324, 1212–1213 (2020).
12. Wang, X. et al. 2017 IEEE Conference on Computer Vision and Pattern Recognition 3462–3471 (Institute of Electrical and Electronics Engineers, 2017).
13. Irvin, J. et al. The Thirty-Third AAAI Conference on Artificial Intelligence 590–597 (Association for the Advancement of Artificial Intelligence, 2019).
14. Johnson, A. E. W. et al. Sci. Data 6, 317 (2019).
15. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. 2017 IEEE Conference on Computer Vision and Pattern Recognition 4700–4708 (Institute of Electrical and Electronics Engineers, 2017).
16. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. Pac. Symp. Biocomput. 26, 232–243 (2021).
17. U.S. Food & Drug Administration. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/postmarket-requirements-devices (2018).
18. Ferryman, K. J. Am. Med. Inform. Assoc. 27, 2016–2019 (2020).

Author contributions

E.W., K.W. and J.Z. designed the study. E.W., K.W. conducted research with help from R.D. and D.O. All the authors contributed to interpretation of the results and writing of the manuscript.

Competing interests

The authors declare no competing interests.