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

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

[1]Department of Electrical Engineering, Stanford University
[2]Department of Biomedical Data Science, Stanford University
[3]Department of Dermatology, Stanford University
[4]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center
[5]Stanford Law School, Stanford University
[6]Chan-Zuckerberg Biohub

Abstract

Medical AI algorithms are increasingly proposed for the assessment and care of patients. However, there are no established best practices for evaluating such algorithms to ensure reliability and safety. In this paper, we provide a systematic analysis of the evaluation process for 130 medical AI devices approved by the Food and Drug Administration (FDA). We find that 126 of the 130 devices were evaluated in retrospective studies, which are less reliable than prospective evaluations. Furthermore, only 28% of the devices reported evaluating their algorithm on more than one clinical site, and less than 13% reported the performance of their algorithm across demographic subgroups. The small number of sites limits the ability to assess the reliability of algorithms across diverse settings and populations. We conduct a case study on pneumothorax triage devices and find that evaluating deep learning models on a single site alone can mask weaknesses in generalizability with a significant degradation in performance across sites with geographic and demographic differences. Our analysis reveals potential limitations in how medical AI is currently evaluated. We provide recommendations and highlight the importance of more rigorous multi-site testing.

Introduction

There is growing interest in deploying AI algorithms in clinical medicine. Researchers have demonstrated the ability of AI to perform tasks such as classifying tumors in mammograms[1–4], detecting lung conditions like pneumonia in CT images[5], predicting risk of stroke in MRIs[6], and assessing heart failure from cardiac ultrasound[7]. Medical AI algorithms have also received criticism for lacking transparency in methodology[8–12], low sample size[13], and the inability to generalize to broader diverse populations[14,15]. The academic community has started to develop reporting guidelines for AI clinical trials[16–18], which include guidelines such as reporting where and how the data is sourced, inclusion and exclusion criteria, and potential algorithmic bias. However, such methodologies have not been universally adopted for medical AI devices.

In the US, the FDA is responsible for approving all commercially marketed medical AI devices. Most devices are approved as a 510(k) device, which requires showing that the device is substantially equivalent to a previously approved device (a predicate). The FDA releases publicly available information on approved devices in the form of a summary document, which generally contains information regarding the device description, indications for use, comparison to the predicate device (for 510(k) devices), and performance data of the device evaluation study.

The path to safe and robust clinical machine learning requires addressing important regulatory questions. Are medical devices able to demonstrate performance that generalizes to the entire intended population? Are commonly faced shortcomings of machine learning (overfitting to training data, distribution shifts, bias against underrepresented patient subgroups) adequately understood and quantified? The FDA has recently called for improving test data quality, better controls around access to test data, improving trust and transparency with users, monitoring of algorithmic performance and bias on the intended population, and testing with clinicians in the loop[19,20]. To understand the extent to which these concerns are addressed in practice, we analyze the clinical performance data of FDA-approved AI medical devices by creating a database based on potential clinical risk and bias. Then, as a case study, we train deep learning models to detect pneumothorax and evaluate their performance across different hospital sites and patient populations. Finally, we discuss the importance of prospective studies and multi-site evaluations in clinical studies.

Methods

Curating a comprehensive database of FDA-approved medical AI. We aggregated all the FDA-approved devices from January 2015 to December 2020 from the FDA online database of 510(k), De Novo, and Premarket Approval (PMA) approved devices[21–23]. Because searching for specific terms on FDA.gov is not possible[24], we downloaded the PDF file for each approved device summary document, extracted the text, and filtered for AI/ML keywords to create our initial corpus. Then, we merged this corpus with two existing databases of FDA-approved AI devices[24,25] and filtered for AI/ML relevance to create a comprehensive database (Figure 1).

Evaluation information. 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 evaluation test data was collected and evaluated concurrently with device deployment (prospective) or the test set was collected prior to device deployment (retrospective); and whether stratified performance by disease subtypes or across demographic subgroups is reported.

Device information. For each device, we also extracted the following metadata features from the FDA summary document: the name of the device; the submission date, approval date, and the FDA-assigned classification product code; cited predicate and reference devices (for 510(k) devices); the body area of the underlying medical condition treated; and the data modality used

by the device. We assessed the risk level (1 to 4) of each device using the FDA proposed guidelines (see Extended Methods).

Results

In total, we compiled 130 approved devices that meet our review criteria (Supp. Table 6). To our knowledge, this is the most comprehensive database of FDA-approved AI medical devices. Table 1 summarizes four examples of AI devices. Figure 2 displays all devices broken down by body area, risk level, prospective/retrospective studies, and multi-site evaluation.

Most evaluations only perform retrospective studies. Most of the AI devices (126 of 130) only perform retrospective studies in their submission. None of the 54 high-risk devices (category 3 or 4) performed prospective studies. The evaluation datasets for the retrospective studies were collected from clinical sites prior to evaluation, and for most devices, the endpoints measured did not involve a side-by-side comparison of clinicians' performances with and without AI.

For instance, DEN170023 (ContaCT), the first approved AI stroke triage device, sourced 300 cases from two clinical sites according to specified inclusion/exclusion criteria. The ground-truth labels were established by a panel of radiologists. The algorithm was then evaluated on this test dataset, and sensitivity and specificity were reported. Another example is DEN180005 (OsteoDetect), a wrist fracture detection device, where 1000 evaluation images were selected from an existing database according to specified inclusion/exclusion criteria. The ground-truth labels were established by three orthopedic surgeons. The OsteoDetect algorithm was then evaluated on the test dataset, and the area under the receiver operating characteristic curve (AUC) along with sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) were reported as endpoints. Neither device approval required prospective testing nor testing on data from more than one site.

Most FDA reports do not state the number of evaluation sites and only roughly half report sample size. Among the 130 devices we analyzed, 93 devices do not publicly report multi-site evaluation as a part of their evaluation study. Of the 41 devices that reported the number of evaluation sites, four devices were evaluated in only one site and eight devices were evaluated in only two sites. This suggests that a substantial proportion of approved devices could have been evaluated only at a small number of sites, which often tend to have limited geographic diversity[15]. The number of approvals for AI devices has increased rapidly in the past five years, with over 75% of approvals coming in the past two years and over 50% in the last year. However, the proportion of approvals with multi-site evaluation and reported sample size has remained relatively stagnant during the same period of time (Supp. Figure 2). Furthermore, the FDA reports of 59 devices (45%) do not state sample size in their studies. Of the 71 devices that do, the mean is 474 and the median is 300, with a standard deviation of 619.

Case study of multi-site evaluation for pneumothorax detection. Since it is critical to understand how a model's performance generalizes to the entire intended population, we

explored how deep learning models might perform when evaluated on populations from multiple clinical sites that represent different populations. To this end, we choose pneumothorax (collapsed lung) detection as a case study since there are currently four 510(k)-cleared medical devices approved for triaging 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 use three datasets, each from a different hospital site in the US: 1) the NIH Clinical Center in Bethesda, MD (abbreviated as NIH)[26], 2) Stanford Health Care Hospital in Stanford, CA (abbreviated as SHC)[27], and Beth Israel Deaconess Medical Center in Boston, MA (abbreviated BIDMC)[28]. NIH, SHC, and BIDMC contain 112120, 223414, and 53848 total images with 5302, 19448, 10358 images positive for pneumothorax, respectively. We use a DenseNet-121 architecture as the deep learning model backbone, which has been demonstrated to be a top-performing model[14,27] for classifying chest conditions. Full training details are available in Extended Methods.

To quantify how the AI's performance varies across sites, we trained three separate models using the patient data from 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. The results are summarized in Table 2; for example, the first row indicates 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. Across the board, we find substantial drop-offs in model accuracy when evaluated on a different site. For example, the NIH-trained model achieves an AUC of 0.883 on its own test data but only has an AUC of 0.759 when tested on BIDMC. Some of the performance variations could be due to differences in patient compositions across sites. For instance, we observe that the performance disparity between White and Black patients in the BIDMC test set increases from 0.024 AUC within BIDMC to 0.043 AUC and 0.109 AUC when evaluated on the other two sites (Supp. Table 2). We also compute AUC when controlling for each site's demographic proportions (for age and sex) and performance differences remain comparable, suggesting that cross-site differences are not solely driven by demographics (Supp. Table 5).

Analyzing the potential limitations of retrospective studies. In settings where devices are evaluated retrospectively, the test dataset can be known before the model is deployed and evaluated. Given that test data leakage is a cited concern of the FDA[19], we sought to quantify its potential effect by measuring how much we can improve the test AUC if given the option of cherry-picking among models for which we know the test performance. Each of the three models (NIH, SHC, BIDMC) was trained five times using the same training dataset but with different random seeds and evaluated individually on the test set. Across the five trained models, we observe an average difference of up to 0.064 AUC using test data subsets of 645 individuals, which simulates the average sample size of the four pneumothorax detection devices (Supp. Table 3 and 4). This shows that it is easy to achieve higher test results even with a minimal amount of peeking at the test set, which is possible with retrospective but harder with prospective evaluation. Test sets with small sample sizes can be especially vulnerable.

Discussion

In the last few years, AI served as the backbone of many state-of-the-art algorithms in healthcare research. While such algorithms have been shown to match or exceed the performance of their non-AI based counterparts, AI is well known to be susceptible to overfitting on training populations leading to poor and/or biased predictions, unclear failure modes, and difficulty in explaining model predictions. Accompanying this advance in research performance has been a wave of efforts to commercialize these algorithms in the form of FDA-approved medical devices. In this paper, we ask whether the regulatory process used to evaluate AI medical devices has evolved to adequately address the unique issues of this new class of algorithms. In particular, we sought to understand how well the current FDA review framework captures the robustness and reliability of the devices across diverse populations.

Related works The FDA does not maintain a curated list of medical devices that use AI/ML. As of December 2020, there are two publicly available databases containing FDA-approved AI medical devices. First, the American College of Radiology[25] provides an updated list of FDA-cleared AI devices pertaining to radiology. Second, Benjamins et al.[24] have aggregated a dataset of AI/ML-based medical devices. Our work presents a more comprehensive list of devices, and include study details and risk and bias assessment, whereas the other two do not.

Previous studies[11,16] have conducted systematic reviews of academic papers and clinical trials of medical AI algorithms but have not studied FDA-approved AI devices. Others[29] reported a decrease in performance for convolutional neural networks (CNNs) on external sites when compared to internal sites, but focused on pneumonia, which is currently not addressed by any FDA-approved devices. One study[30] evaluated mammography AI algorithms on an external site but does not provide comparisons to baseline performances. Related to our pneumothorax case study, a previous study[14] reported performance disparities across demographic subgroups with CNNs trained on multiple chest diseases but did not report cross-site performance.

Multi-site evaluation Most clinical studies do not report multi-site performance or only evaluate on one site, which can result in significant performance disparities when generalized to other sites. While the number of sites used in a study is available to the FDA, it is also important to consistently report it in the public summary document in order for clinicians, researchers, and patients to make more informed judgments about the reliability of the algorithm. In our case study on pneumothorax detection triage devices, we find that while the within-site performance remains high (average AUC of 0.893), the performance significantly degrades by an average of 0.072 AUC and up to 0.124 AUC when evaluated on the other two sites. We also observe that performance disparities between White and Black patients are exacerbated in the other two sites. Multi-site evaluations are important for understanding algorithmic overfitting and bias and can help account for variations such as the equipment used, technician standards, image storage formats, demographic makeup, and disease prevalence.

Retrospective studies Almost all FDA-approved medical AI devices were evaluated retrospectively. Since retrospective studies assess the performance of the algorithm only after

clinical action and outcome has already taken place, they may not provide a direct comparison with the existing standard of care and are more prone to overfitting or bias in data acquisition.

Whether the data is collected retrospectively or prospectively, bias can be introduced if there is any information leakage that occurs from the test set. This leakage may occur more frequently with retrospective studies as early access to any of the test data, even unlabeled data, can allow medical device applicants to extract statistics from images and fine-tune on similar training data. With either retrospective or prospective data collection, medical device applicants might perform preliminary evaluations, or peeks, on the test population. Even a limited number of peeks, however, can allow for models to be cherry-picked over a set of hyperparameters. From our case study, the 0.064 AUC difference across the five models with different random seeds would be sufficient to drop the test AUC of all four devices below 0.95, the minimum bar set by the FDA for these devices[30]. Prospectively evaluating algorithms, with clearly delineated guidelines on access to the test population data, guards against gaming and overfitting to test datasets.

With comparison to standard of care, prospective studies can also more properly assess a device's effectiveness when deployed in clinical settings and benefit from randomization to truly understand the effect of an intervention. Prospective studies facilitate rigorous evaluation of the impact of the AI decision tool on clinical practice, which is important given that human-computer interaction can deviate substantially from a model's intended use. For example, most computer-aided detection (CAD) diagnostic devices are intended to be decision support tools rather than primary diagnostic tools. A prospective randomized controlled study may reveal that clinicians are misusing this tool as a primary diagnostic tool and that outcomes are different than what would be expected if the tool was used as decision support.

Choice of predicate devices Predicate devices establish regulatory safeguards around classes of AI approvals, such as diagnostic or triage tools. For instance, ContaCT (DEN170073), the first AI triage device approved by the FDA, is used for finding cases suggestive of large vessel occlusion (stroke) in the brain. Devices do not need to be approved for the same condition as its cited predicate: 12 of the 24 of ContaCT's subsequent 510(k) devices are approved for non-head related conditions (Supp. Figure 1). Importantly, 510(k) devices must demonstrate that it is as safe and effective as the predicate[23]. However, our findings indicate that important study protocols like multi-site evaluation and reporting demographic data are missing in many subsequent 510(k) devices. Among the top-five most cited predicates with multi-site evaluation, only an average of 47% of subsequent devices reported multi-site evaluation. Similarly, among the top-five most cited predicates with demographic data, only 32% of subsequent devices report demographic data. Consistent regulatory standards for clinical studies are important for ensuring that approved devices are uniformly safe and effective against bias and overfitting.

Second, ML-specific performance defects may be overlooked due to an inadequate regulatory standard set by non-ML predicate devices. For instance, several approved medical image enhancement devices (K191688, K183046, K182336) rely on convolutional neural networks to

improve image quality and reduce noise. However, these algorithms are known to introduce visual artifacts or remove important structural information[31], which may cause harm to patients if the effect is not properly understood and addressed. Of note, none of these three devices mention visual degradation as a potential risk, and they all cite predicates that do not use ML algorithms. ML devices without ML-based predicates should include additional special controls on regulating ML-specific performance concerns. Establishing rigorous and appropriate predicate devices can help set a regulatory standard for curbing undesirable performance defects in subsequent AI devices.

Recommendations. Evaluating the performance of AI devices in multiple clinical sites is important for ensuring that algorithms perform well across representative populations and minimize bias. Encouraging prospective studies with comparison to standard of care reduces the possibility of harmful overfitting and more accurately measures true clinical outcomes. Post-market surveillance on AI devices is also needed to understand and measure unintended outcomes and biases that are not detected in prospective, multi-center trials[32,33]. Requiring applicants to use clinical research organizations (CROs) or third-party data escrow services to store test data and perform evaluations can mitigate improper test data access.

Limitations. Risk level categorization has a high degree of variability and subjectivity. In order to make our risk determination as objective as possible, we obtained risk scores for each device from two board-certified doctors. A consensus was reached for the minority of devices with different scores. Additionally, the data obtained only captures the information reported in the summary document made publicly available by the FDA and does not necessarily imply that no such data was reported to the FDA.

Extended methods

Aggregating all device documents from FDA.gov

To provide a systematic and repeatable corpus of AI-related FDA-approved medical devices, we created an automated pipeline for downloading all FDA approvals available on FDA.gov (~18.5k total). We extract the raw text from the metadata and summary documents of all 510(k), De Novo, and PMA devices from January 2015 to December 2020 and filtered for the following keywords: [*deep learning, machine learning, neural network, artificial intelligence].* In total, we found 109 devices that matched at least one of the keywords. We performed a manual review of each device for false positives and found 5 that only contain spurious mentions (eg. *'does not use artificial intelligence'*), leaving 104 approvals. No PMA devices fit our criteria.

Filtering devices based on whether machine learning was used

Since the goal of our analysis is to study modern machine learning, which may be more prone to overfitting and distribution shifts, we conducted an additional review of products that do not fall within this scope. For the purposes of this analysis, we define AI/ML to be algorithms that optimize an objective function via training data. We do not require that the models are trained in a supervised manner, or that the models are deep neural networks. We encountered 53

products that were marketed as artificial intelligence but do not fit this definition as they represent expert or logic-based systems.

Merging with existing datasets
Of the two existing datasets of FDA-approved AI medical devices, Benjamins et al. contains 66 devices, while the American College of Radiology et al. contains 79 devices. The union of these two datasets yields 123 total devices, which we filtered down to 70 based on our criteria for AI/ML. Finally, we combine this with our dataset of 104 devices to yield a sum total of 130 devices.

Risk level assessment
The FDA proposed[34] categorizing devices using the International Medical Device Regulators Forum (IMDRF) risk categorization framework, which assigns a 1 to 4 risk level to each Software as a Medical Device (SaMD) according to 1) the significance of the information provided by the SaMD to the healthcare decision, and 2) the state of healthcare situation or condition. The risk assessment calculus is shown in Supp. Table 1.

In determining significance, we used the following classification product codes to identify diagnostic products: QBS, QDQ, POK, OEB, PIB, DPS. To identify devices that drive clinical management, we filtered for triage products and matched devices that mentioned “triage and notification”. Additionally, we include devices that produce information used toward clinical decision-making, such as breast density tools or lesion segmentation. We assign risk based on the significance of the ML-specific component of the device in cases where the device performs other functions (ie. an image visualization software that includes an automated segmentation tool). Since the severity of medical conditions relies on clinical interpretation and expertise, we obtained categorizations from two board-certified physicians independently and then formed consensus decisions on devices under disagreement. According to IMDRF, non-serious conditions are those that generally do not yield long-term irreversible consequences on patients’ health (e.g. astigmatism, cystic acne). Serious conditions include diseases that are moderate in progression and in which treatment is not generally considered time-sensitive (e.g. heart arrhythmia, diabetes). Critical conditions are generally life-threatening or irreversibly damaging (e.g. stroke, meningitis).

Summary document details
We extract the following features from the summary documents available on the FDA website. (Supp. Table 6; NA indicates that a value was not reported):

- FDA-assigned approval number for the device (*approval_number*). 510(k) devices start with ‘K’ and De Novo devices start with ‘DEN’.
- Name of FDA-approved device (*device_name*).
- Classification product code used to identify the generic category of a device for FDA (*classification_product_code*).
- Date that the device submission was received by the FDA (*date_received*).
- Date of FDA approval for the device (*decision_date*).

- Indications for use for the device (*indications_for_use*).
- Body area of the underlying condition or disease treated (*body_area*).
- Modality used to acquire the images or data, eg. PET, MRI, X-ray (*modality*).
- For 510(k) devices, the predicate device(s) cited, along with other reference devices (*predicate_numbers*).
- Number of patients enrolled in the evaluation study, if available (*sample_size*). If there were multiple studies conducted, we reported the highest sample size.
- Number of distinct clinical sites that the test population is sourced from (*num_sites*).
- Whether the study specified the inclusion and/or exclusion criteria for the evaluation study (*inc_exc_criteria*).
- Whether the study explicitly mentions data that is sourced across demographic subgroups, ie. sex, age, race, socioeconomic status (*has_demographics*).
- Whether the study was performed prospectively or retrospectively (*is_prospective*).
- Whether the study stratified performance by disease subtypes (*subtype_analysis*).
- Significance of the information provided by the device to the healthcare decision (0 - inform clinical management, 1 - drive clinical management, or 2 - treat or diagnose) (*significance*).
- State of healthcare situation or condition that the device treats (0 - non-serious, 1 - serious, or 2 - critical) (*severity*).
- Overall risk level of the device (using Supp. Table 1) (*risk_level*).
- AI/ML keywords that appeared in the summary document (*keywords*).
- URL of the summary document PDF (*summary_link*).

Deep learning model training details

On each dataset, we train five models with identical hyperparameters, except for randomness in data augmentations and training data ordering. An initial learning rate of 1e-5 is used with Adam[35] optimizer and a batch size of 16. The learning rate is halved if the validation AUC has not improved in 50k steps, and training stops if it has not improved in 150k steps. Images are scaled to 512x512 pixels with zero-padding on the short side. The data is randomly augmented using up to 15 degrees rotation, 5% translation, 5% scaling, all sampled uniformly. Predictions are calculated by averaging scores across the five model outputs per image, and then averaging scores across images in a study. The study level AUC is reported, with the 95% confidence interval calculated using bootstrap sampling of the test set. Each dataset is split randomly into 80% training, 10% validation, and 10% test subsets. Across all datasets, NaN labels are treated as negative. All models are pre-trained on ImageNet[36], with images scaled to a [0,1] value range and normalized on ImageNet pixel statistics. Positive and negative examples are sampled approximately equally.

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.

Acknowledgements. J.Z. is supported by NSF CCF 1763191, NSF CAREER 1942926, NIH P30AG059307, NIH U01MH098953 and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

Code and Data availability. Our compiled database will be available as a downloadable CSV file. All code used to produce the database will be made available via Jupyter Notebooks. The code used to train the pneumothorax detection deep learning models along with model weights will also be available on Github. Full resources are accessible at: *https://github.com/ericwu09/medical-ai-evaluation*

Figures and tables

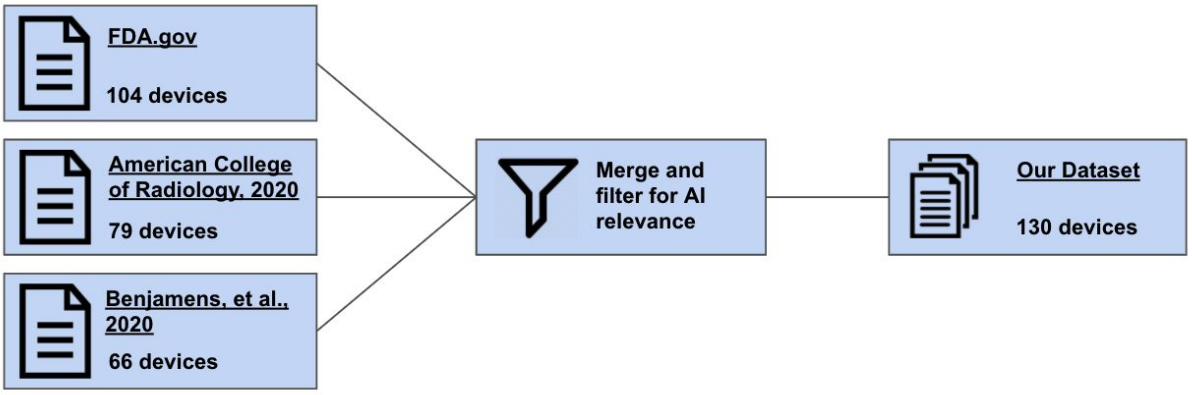


Figure 1. Curation of FDA-approved medical AI algorithms. We aggregated all of the approved AI devices from FDA.gov as of December 4, 2020. We then merged this set with two recent databases of FDA-approved algorithms and removed devices that did not use AI algorithms. The final database contains 130 FDA-approved devices.

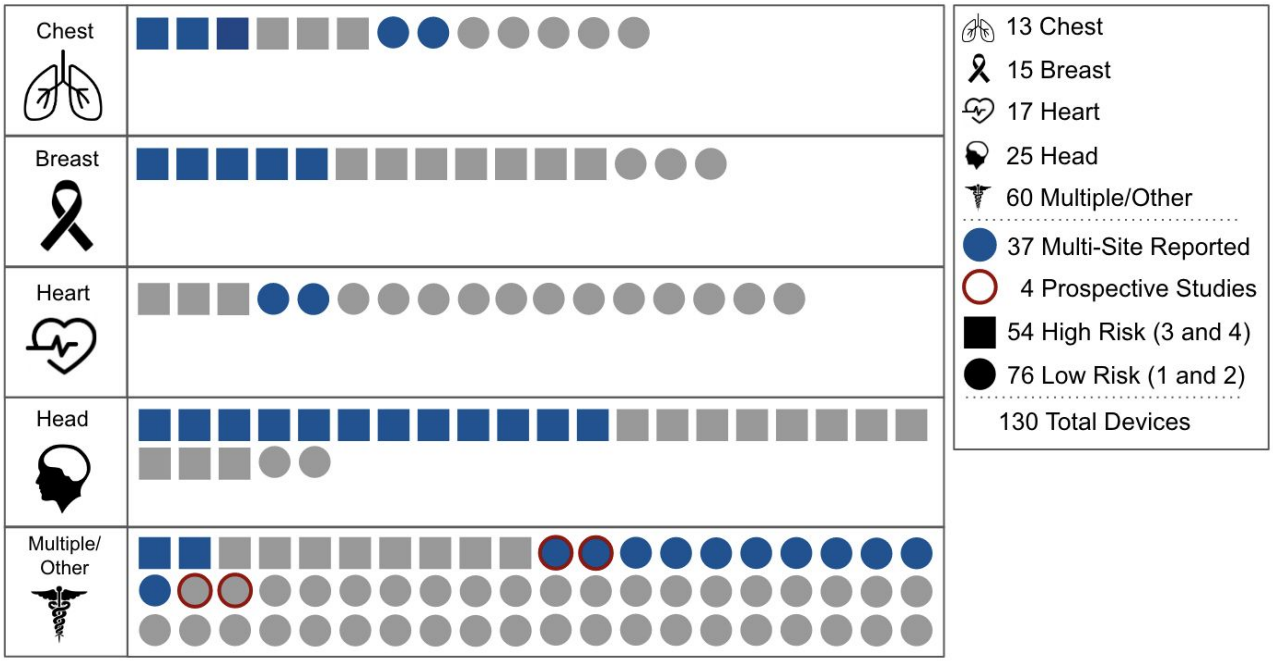


Figure 2. Breakdown of 130 FDA-approved medical AI devices by body area. Devices are categorized by risk levels, where squares indicate high risk and circles indicate low risk. Dark blue indicates a multi-site evaluation was reported, and a red border indicates a prospective study was performed.

Risk level	Body Area	Approval ID	Product Name	Indications for Use	Evaluation sample size	Number of evaluation sites	Includes demographic data
1	Blood	K201301	X100 with Full Field Peripheral Blood Smear (PBS) Application	Automatically locates classifies, and presents images of blood cells on peripheral blood smears	30	3	No
2	Chest	K192320	HealthCXR	Triages chest X-ray cases for pleural effusion	554	N/A	No
3	Head	DEN180005	OsteoDetect	Diagnosis of distal radius fractures in X-ray images of the wrist	200	N/A	Yes
4	Breast	K181704	Transpara	Diagnosis of breast cancer in mammograms	240	2	No

Table 1. Four examples of FDA-approved medical AI devices. N/A indicates that the information is not provided in the public FDA summary document associated with the device.

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**

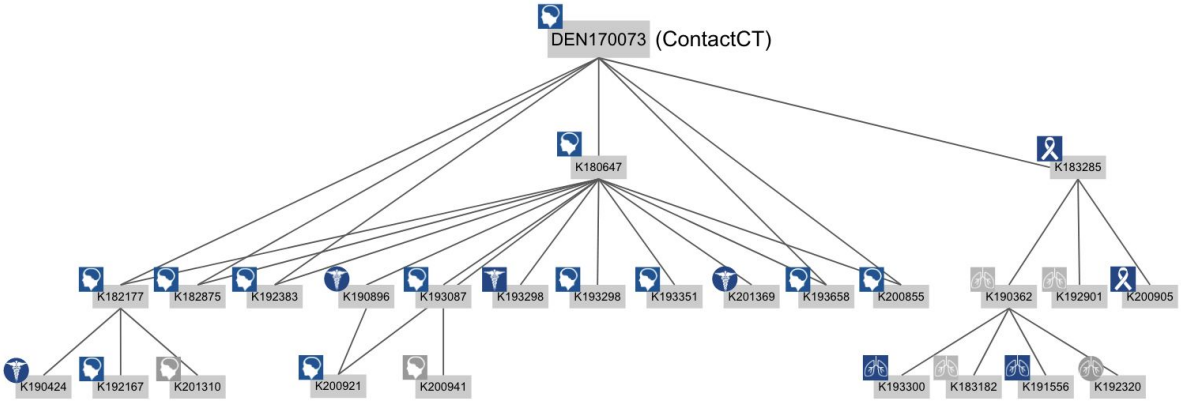
Table 2. Cross-site performance. Each row represents an algorithm trained on a single site, with the columns indicating the dataset the algorithm is evaluated on. Each cell contains the AUC and 95% confidence interval. Bolded numbers indicate within-site performance. For example, the algorithm trained on the NIH data achieves good performance on held-out NIH test patients (AUC 0.883) but performs substantially worse on the BIDMC test patients (AUC 0.759) and SCH test patients (AUC 0.779). The data size (N) refers to the test dataset.

References

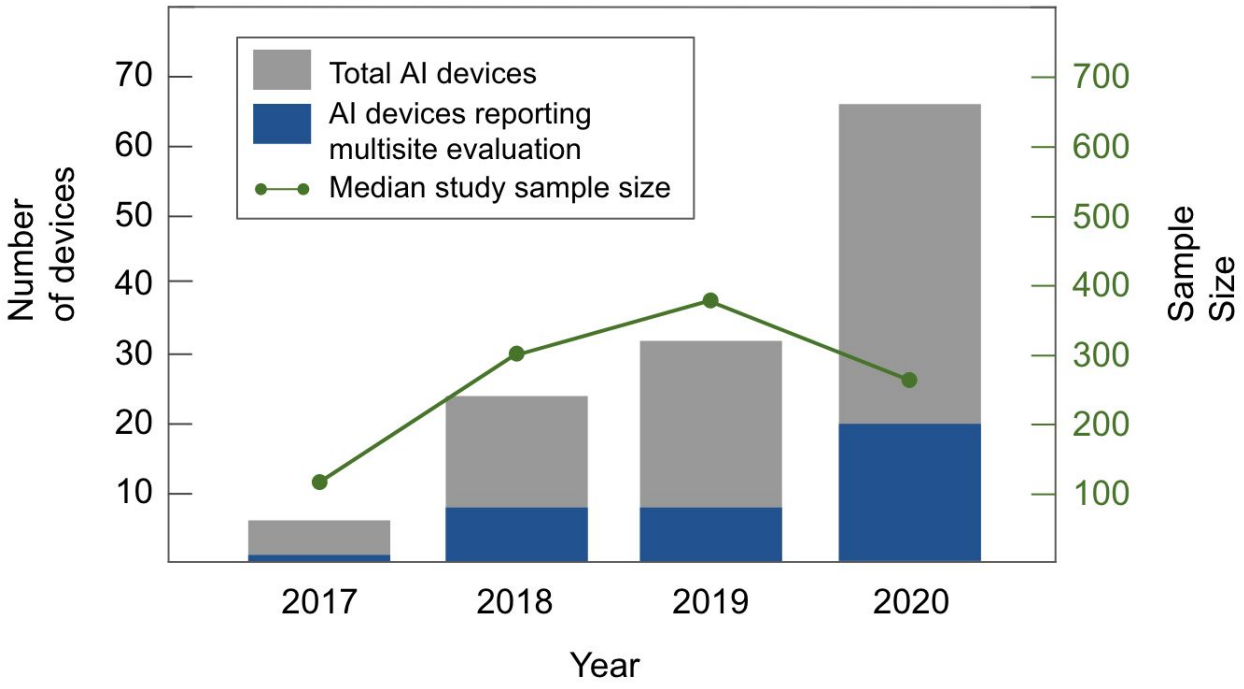
1. Yala, A., Lehman, C., Schuster, T. & Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction • Content code. *Radiology* **292**, 60–66 (2019).
2. Lotter, W. *et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using annotation-efficient deep learning approach*. *arxiv.org* https://arxiv.org/abs/1912.11027 (2019).
3. McKinney, S., Sieniek, M., Godbole, V., Nature, J. G.- & 2020, undefined. International evaluation of an AI system for breast cancer screening. *nature.com*.
4. Ribli, D. *et al.* Detecting and classifying lesions in mammograms with deep learning. *nature.com*.
5. Rajpurkar, P. *et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*.
6. Nielsen, A., Hansen, M. B., Tietze, A. & Mouridsen, K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. *Stroke* **49**, 1394–1401 (2018).
7. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
8. Singh, A., Sengupta, S. & Lakshminarayanan, V. *Explainable deep learning models in medical image analysis*.
9. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. *Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis*.
10. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
11. Nagendran, M. *et al.* Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *The BMJ* **368**, (2020).
12. Haibe-Kains, B. *et al.* The importance of transparency and reproducibility in artificial intelligence research. *arXiv* **10**, 34–34 (2020).
13. Balachandar, N., Chang, K., Kalpathy-Cramer, J. & Rubin, D. L. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *J. Am. Med. Inform. Assoc.* **27**, 700–708 (2020).
14. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *ArXiv200300827 Cs Eess Stat* (2020).
15. Kaushal, A., Altman, R. & Langlotz, C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA - J. Am. Med. Assoc.* **324**, 1212–1213 (2020).
16. 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).
17. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).
18. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension | Nature Medicine. https://www.nature.com/articles/s41591-020-1037-7.
19. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper and Request for Feedback*. https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm514737.pdf.

20. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.* https://www.fda.gov/media/145022/download.
21. Device Classification Under Section 513(f)(2)(De Novo). https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm.
22. Premarket Approval (PMA). https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm.
23. 510(k) Premarket Notification. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm.
24. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 1–8 (2020).
25. FDA Cleared AI Algorithms | American College of Radiology. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms.
26. Wang, X. *et al.* ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* vols 2017-January 3462–3471 (Institute of Electrical and Electronics Engineers Inc., 2017).
27. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *33rd AAAI Conf. Artif. Intell. AAAI 2019 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019* 590–597 (2019).
28. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
29. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**, e1002683–e1002683 (2018).
30. Salim, M. *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol.* (2020) doi:10.1001/jamaoncol.2020.3321.
31. Cohen, J. P., Luck, M. & Honari, S. Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.* **11070 LNCS**, 529–536 (2018).
32. Health, C. for D. and R. Postmarket Requirements (Devices). *FDA* https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/postmarket-requirements-devices (2018).
33. Ferryman, K. Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. *J. Am. Med. Inform. Assoc.* **27**, 2016–2019 (2020).
34. US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.
35. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).
36. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *ArXiv14090575 Cs* (2015).

Supplemental Figures and Tables



Supp. Figure 1. Illustration of ContaCT (DEN170073) and its relationship to devices that cited it as a predicate device. The icons indicate the body part the device treats (same as in Table 2); squares indicate high risk and circles indicate low risk; and dark blue indicate a multi-site evaluation was reported. There is substantial variation among the devices in terms of risk, body area treated, and multi-site evaluation reporting though they all share the same predicate ancestor.



Supp. Figure 2. The number of approved devices by year, number of approved devices that report multi-site evaluation, and the median evaluation study sample size by year (right y-axis in green). Despite an increased number of device approvals in recent years, the median sample size and relative proportion of devices reporting multi-site evaluation remain stagnant.

State of healthcare situation or condition	Significance of information provided by SaMD to healthcare decision		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Supp. Table 1. The International Medical Device Regulators Forum (IMDRF) framework for risk classification as proposed by the FDA in a recent discussion paper on regulating AI/ML devices[34]. A risk score is assigned to each device according to the seriousness of the underlying treated condition and the contribution of information from the AI device toward clinical decision making.

Patient Information	**Evaluation AUC on BIDMC**		
Model Training Site	BIDMC	SHC	NIH
Sex			
Male	0.888 ± 0.013	0.877 ± 0.015	0.771 ± 0.020
Female	0.889 ± 0.015	0.858 ± 0.022	0.753 ± 0.025
Age			
0-20	0.995 ± 0.013	0.984 ± 0.026	0.771 ± 0.325
20-40	0.877 ± 0.027	0.895 ± 0.027	0.821 ± 0.038
40-60	0.894 ± 0.015	0.884 ± 0.020	0.791 ± 0.025
60-80	0.889 ± 0.015	0.860 ± 0.021	0.748 ± 0.028
80+	0.889 ± 0.015	0.825 ± 0.045	0.697 ± 0.050
Race/Ethnicity			
White	0.883 ± 0.011	0.871 ± 0.015	0.759 ± 0.020
Black	0.907 ± 0.023	0.828 ± 0.046	0.650 ± 0.066
Hispanic/Latino	0.883 ± 0.058	0.868 ± 0.069	0.755 ± 0.091
Asian	0.892 ± 0.033	0.891 ± 0.040	0.834 ± 0.054
American Indian/Alaska Native	0.936 ± 0.056	0.936 ± 0.086	0.641 ± 0.168
Other	0.872 ± 0.033	0.857 ± 0.039	0.800 ± 0.036
Insurance			
Medicaid	0.902 ± 0.028	0.883 ± 0.033	0.812 ± 0.046
Other	0.888 ± 0.010	0.871 ± 0.013	0.763 ± 0.016

Supp. Table 2. The performance of the models across 15 subgroups of BIDMC patients stratified by age, sex, race, and whether the patient used Medicaid insurance (a proxy for socioeconomic status). The column name indicates the dataset for which the model was trained on. Demographic subgroup discrepancies may be more apparent in multi-site evaluations. For instance, the difference in performance in White and Black patients increases from 0.024 AUC with BIDMC to 0.043 AUC and 0.109 AUC when evaluated using the other two models.

Approval Number	Device Name	Sample Size	AUC	95% CI
K183182	Critical Care Suite	804	0.960	0.949, 0.972
K190362	HealthPNX	588	0.983	0.974, 0.9902
K193300	AIMI-Triage CXR PTX	300	0.967	0.950, 0.984
K191556	Behold.ai red dot	888	0.975	0.966, 0.984

Supp. Table 3. Summary of the evaluation studies for the four pneumothorax triage devices approved by the FDA. Data is extracted from the FDA summary documents for each device.
The higher observed AUCs from the device evaluation studies compared to our case study could be explained by the device algorithms having enriched test datasets, more precise ground-truth labeling, and training on larger in-house datasets. We also emphasize that the relative AUC differences, rather than absolute AUCs, are the focal point for elucidating multi-site performance disparities.

Site	**SHC**	**BIDMC**	**NIH**
SHC	**0.905 ± 0.048**	0.870 ± 0.078	0.854 ± 0.082
BIDMC	0.825 ± 0.067	**0.893 ± 0.060**	0.838 ± 0.090
NIH	0.782 ± 0.072	0.767 ± 0.094	**0.885 ± 0.062**

Supp. Table 4. We report the results as in Table 2, except using a sample size of N=645 (the average sample size across four pneumothorax devices) when computing the confidence intervals with bootstrapping. We observe that the average one-sided 95% CI increases from 0.014 in Table 2 to 0.073 shown here. This suggests that larger evaluation test sets increase the reliability of the reported evaluation results.

Site	**SHC**	**BIDMC**	**NIH**
SHC	**0.897 ± 0.009**	0.871 ± 0.013	0.845 ± 0.023
BIDMC	0.825 ± 0.011	**0.890 ± 0.003**	0.825 ± 0.022
NIH	0.774 ± 0.013	0.782 ± 0.014	**0.883 ± 0.015**

Supp. Table 5: We control for the effect of demographic distribution in each site by matching the testing site for each algorithm to the site it was trained on. Each row represents the AUC of an algorithm tested on each site with bootstrap sampling to match the age and sex of the training set, with the 95% confidence intervals. We observe that the results are largely within the original 95% confidence intervals, suggesting that performance differences between sites cannot be explained by demographics alone.

Supp. Table 6: Full database of all 130 FDA-approved medical AI devices, included in the attached supplementary file. Columns are described in Extended Methods.