Regulating AI Adaptation: An Analysis of AI Medical Device Updates

Kevin Wu[*]   KEVINYWU@STANFORD.EDU
Stanford University, USA

Eric Wu[*]   WUE@STANFORD.EDU
Stanford University, USA

Kit Rodolfa   KRODOLFA@LAW.STANFORD.EDU
Stanford University, USA

Daniel E. Ho[†]   DHO@LAW.STANFORD.EDU
Stanford University, USA

James Zou[†]   JAMESZ@STANFORD.EDU
Stanford University, USA

Abstract

While the pace of development of AI has rapidly progressed in recent years, the implementation of safe and effective regulatory frameworks has lagged behind. In particular, the adaptive nature of AI models presents unique challenges to regulators as updating a model can improve its performance but also introduce safety risks. In the US, the Food and Drug Administration (FDA) has been a forerunner in regulating and approving hundreds of AI medical devices. To better understand how AI is updated and its regulatory considerations, we systematically analyze the frequency and nature of updates in FDA-approved AI medical devices. We find that less than 2% of all devices report having been updated by being re-trained on new data. Meanwhile, nearly a quarter of devices report updates in the form of new functionality and marketing claims. As an illustrative case study, we analyze pneumothorax detection models and find that while model performance can degrade by as much as 0.18 AUC when evaluated on new sites, re-training on site-specific data can mitigate this performance drop, recovering up to 0.23 AUC. However, we also observed significant degradation on the original site after re-training using data from new sites, providing insight from one example that challenges the current one-model-fits-all approach to regulatory approvals. Our analysis provides an in-depth

look at the current state of FDA-approved AI device updates and insights for future regulatory policies toward model updating and adaptive AI.

Data and Code Availability The primary data used in this study are publicly available through the FDA website. Our analysis of the data and code used is publicly available on GitHub: https://github.com/kevinwu23/AIUpdating.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

While the number of AI products developed for commercial applications is rapidly growing, the implementation of robust regulatory frameworks still lags behind (Larson et al., 2021; Wirtz et al., 2020; Wu et al., 2021a). Recently, high-profile accidents involving Boeing (Wendel, 2019) and Tesla (Corfield et al., 2023) have been attributed to issues with software and AI updates in their systems. Applications of AI to consumer lending (Johnson et al., 2019) and hiring systems (Bogen and Rieke, 2018) has also led to calls for more flexible regulatory systems that can anticipate algorithmic changes and biases. Such cases highlight the inherent challenges regulators face due to the adaptive nature of software and especially AI products: while model adaptation and updates are a necessary step in maintaining or improving their

---

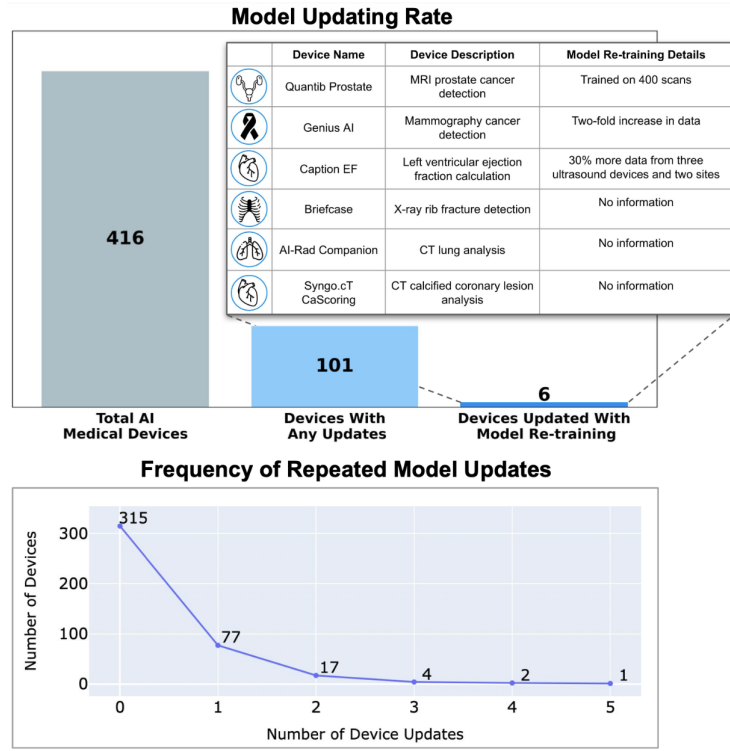[*] Equal first authorship

[†] Equal senior authorship

Model Updating Rate

	Device Name	Device Description	Model Re-training Details
	Quantib Prostate	MRI prostate cancer detection	Trained on 400 scans
	Genius AI	Mammography cancer detection	Two-fold increase in data
	Caption EF	Left ventricular ejection fraction calculation	30% more data from three ultrasound devices and two sites
	Briefcase	X-ray rib fracture detection	No information
	AI-Rad Companion	CT lung analysis	No information
	Syngo.cT CaScoring	CT calcified coronary lesion analysis	No information

416

101

6

Total AI Medical Devices	Devices With Any Updates	Devices Updated With Model Re-training

Frequency of Repeated Model Updates

Figure 1: (Top) Proportion of devices that report model re-training and other update types. The devices with any updates are a subset of the total AI medical devices, and the devices updated with model re-training are a subset of devices with any updates. The call-out provides details on the six devices that received model re-training, along with device name, device description, and details provided within the FDA approval regarding the type of model re-training applied. (Bottom) Graph of the number of times devices have been updated. The x-axis refers to the number of successive updates, and the y-axis refers to the count of devices in each group.

performance, they can also introduce unknown safety risks (Babic et al., 2019; Gilbert et al., 2021).

In the US, the Food and Drug Administration (FDA) has been an early mover in AI regulation, with over 500 approved submissions for AI devices as of 2022 (Center for Devices and Radiological Health, 2022). The FDA faces unique challenges with regard to model updating, as adverse events can directly compromise patient well-being. As such, the FDA has traditionally not allowed any changes to a model once it has been approved (Gerke et al., 2020). At the same time, AI models are well-known to be prone to distribution shifts, whereby variations in factors such as medical practice, patient demographics, or disease prevalence can significantly affect a model's performance (Raghu et al., 2019; Wiens et al., 2019; Wong et al., 2021). For example, researchers recently found that Epic's widely used sepsis prediction model performed much worse than initially reported after being deployed in a new hospital setting (Wong et al., 2021). Such cases demonstrate that fixed AI models that never receive updates can likewise compromise patient safety. Recently, the FDA has taken action to address the limitations of a fixed-model approach by providing guidelines for a potential Predetermined Change Control Plan (PCCP) (Center for Devices and Radiological Health, 2023a), as well as a document describing best practices in machine learning published jointly by US, Canadian, and UK health authorities (Center for Devices and Radiological Health, 2023c). Under this provision, developers can make a limited set of changes to their models

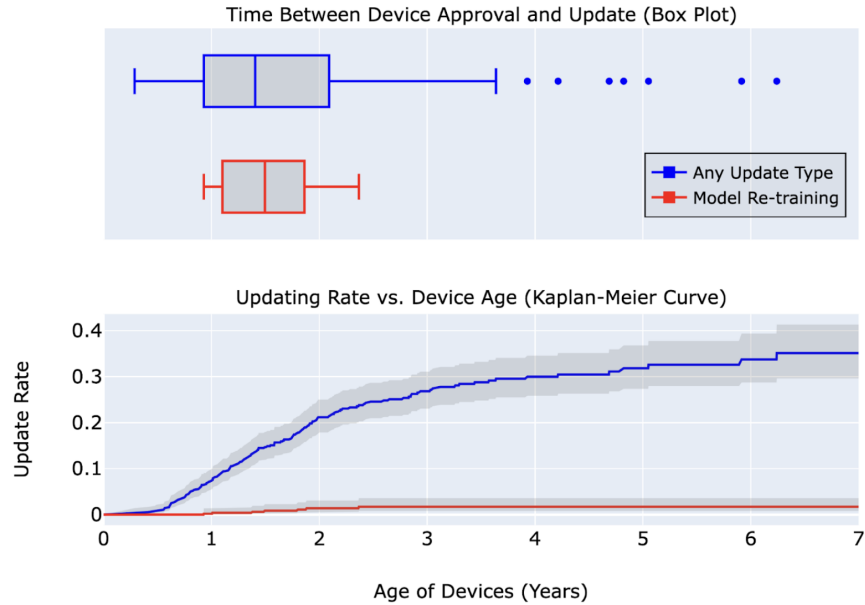Time Between Device Approval and Update (Box Plot)

Figure 2: (Top) Time to device update (in years), represented as a box plot where each vertical bar represents the 25th and 75th percentile, respectively. The red line represents the distribution of time to update for model re-training, while the blue line refers to time to update of any type. (Bottom) The estimated rate of updating as a function of device age, which is estimated through the Kaplan-Meier function in order to account for the right censorship in the dataset. For example, devices at two years old are updated across all types approximately 20% of the time, whereas the same-age devices are updated with model re-training around 1.4% of the time. Both plots have been cut-off at seven years as the longest device update observed is 5.9 years. Shading indicates the 95% confidence interval around the updating rate at any given time.

without a new submission as long as it is pre-specified in their initial approval. Such proposed measures by the FDA underscore the importance of the ongoing discussion around the appropriate levels of regulation regarding the adaptive nature of AI.

Despite the importance of model updating in AI medical devices, little is known about how often such devices are currently being updated. While AI developers may individually publish press releases about changes to their model, there does not exist a systematic analysis of updating across all AI medical devices. FDA approvals by the same developer often contain variants of company and product names, making it difficult to automatically link devices together. Furthermore, devices under the same name often vary widely according to their use cases and are actually different products. In this study, we aim to resolve these issues by organizing and grouping FDA-approved AI medical devices by their updates and

performing an analysis of the frequency and nature of model updates. Our study explores the extent to which developers choose to update their devices given current regulatory, economic, and technological factors. Additionally, we perform an illustrative case study on AI models designed to predict pneumothorax, evaluating whether model updates consistently yield improved performance when re-trained on target populations.

2. Methods

2.1. Collecting device updates

The primary data for this study consists of FDA approval documents for AI medical devices, which are publicly available through the FDA's online database (www.fda.gov). Under the FDA's 510(k) approval process, developers must demonstrate that the med-

ical device they are marketing (the "subject" device) is "substantially equivalent" to a device already available on the market (the "predicate" device) (Brindza, 1980). Furthermore, each FDA-approved device is classified using a product code that indicates the overall function and safety profile (Center for Devices and Radiological Health, 2023a). For example, the product code QFM refers to "Radiological Computer-Assisted Prioritization Software For Lesions" and includes many common triage-based AI detection software. In our analysis, an FDA approval is considered a device update if 1) the predicate and the subject devices are from the same manufacturer, 2) both devices share the same product classification code, and 3) both devices are AI devices.

When grouping by manufacturer names, multiple variants of the same manufacturer often appear (e.g., Siemens Medical Solutions USA Inc. and Siemens Medical Solutions, Inc.). To reconcile these differences, we first applied approximate string matching with Levenshtein distance and a similarity threshold of 0.8 to create candidate company name groupings before manual review. Furthermore, to systematically identify the predicate devices for each FDA approval, we extracted the PDF texts and performed a search over the first appearance of a submission number outside of the subject device number before performing a manual review.
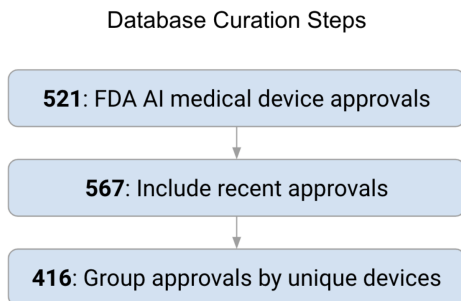
Database Curation Steps

Figure 3: Schematic of database curation steps, starting from the FDA's official list of AI medical device approvals to the final set of 416 unique devices.

Our dataset starts with the FDA's list of AI/ML medical devices, which contains a total of 521 FDA approvals (recent as of 10/5/22). In order to include more recent updates, we added an additional 46 approvals from 10/5/22-07/01/23 that reference one of the 521 approvals as a predicate device. In total, our final dataset contains 416 unique devices, which are represented across 567 total FDA approvals (e.g. a single device can be approved multiple times for each update). The data curation steps and sample sizes are outlined in Figure 3.

After identifying all device updates, we determined the types of updates that occur. For each FDA approval, manufacturers are required to provide details of the subject device's technological comparison to the predicate device. For example, FDA approval K221727 (syngo.CT Extended Functionality) includes a section titled "Comparison of Technological Characteristics with the Predicate Device", which contains a table comparing and contrasting the predicate (SOMARIS/8 VB60) with the subject device (SOMARIS/8 VB70). Within this section, the update is described to have "Improved quality of the bone removal algorithm for the head & neck region", and notes that "Segmentation of the bones use a deep learning algorithm instead of a traditional image processing". We annotated each updated device according to the type of update received, which is further detailed in the Results section.

2.2. Case study

Given that site-specific re-training is not allowed under current FDA 510(k) guidelines, we conducted a case study on pneumothorax detection models for chest X-rays to understand the potential performance gains that are currently uncaptured. There are currently four FDA-approved medical devices for the triage of X-ray images for the presence of pneumothorax (Wu et al., 2021a), 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) (Wang et al., 2017); Stanford Health Care in Palo Alto, California (SHC) (Irvin et al., 2019); and Beth Israel Deaconess Medical Center in Boston, Massachusetts (BID) (Johnson et al., 2023). We used a DenseNet-121 deep-learning architecture (Huang et al., 2017) that has been demonstrated to be a top-performing model for the classification of chest conditions (Irvin et al., 2019; Seyyed-Kalantari et al., 2020). These datasets represent a diversity of patient populations, imaging manufacturers, and pathology reporting standards (Wu et al., 2021b). To quantify how the

Update Example	Example FDA Approval	Example Product	Example Change
Model design	K201310	Caption Health	Simplification of network architecture via simple pooling layers and quantization
Input signal	K211541	Mammoscreen	Added 3D (DBT) mammogram support
Target population	K210034	EnsoSleep	Pediatric Patient Population Included
Training data	K221240	BriefCase	Larger training dataset
Device User	K203508	BriefCase	From radiologist only to medical professionals
Compatibility	K191994	ProFound AI Software V2.1	Added support for Siemens machines
UI/UX	K221240	BriefCase for IHC Triage	New desktop application
Hardware	K221147	Vivid T8, Vivid T9	Modified transducers
Software	K220590	aPROMISE X	Accessible through cloud
Device Use Case	K193417	FractureDetect (FX)	Expanded range of fracture detection from wrist to ankle, elbow, etc.

Table 1: Examples of updating types present in follow-up devices. The table provides the update type, along with an example of each subtype.

AI's performance varies across sites, we trained 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. Similar to top-performing model approaches (Irvin et al., 2019; Seyyed-Kalantari et al., 2020), we trained five identical models (with different random seeds) for each setting and then ensembled the predictions by averaging the predicted probabilities across each model. We then re-trained the model (by fine-tuning) on a small subset of training data of five thousand examples from an unseen external site and re-evaluated the model's performance on both the original and external sites. We perform fine-tuning with the standard approach of updating all the model weights for a fixed number of steps without changing the hyperparameters.

3. Results

3.1. Device Update Frequency and Types

Among our dataset of 416 unique devices, we found that 101 devices report having been updated at least once (Figure 1). However, the vast majority of these updates expand the functionality or marketing claims of the device, essentially constituting a new device rather than a true model update. Of these 101 devices, only six of the updated devices report retraining in the model with new data. For each of the six devices, details on the types of data used in retraining are limited, with only three providing how much training data was used. For AI devices, retraining on new data is central to and distinctive of the technology, leading to our focus on the novel regulatory issues here. For example, Syngo.CT CaScoring (K221219), which analyzes calcified coronary lesions, only references that "the algorithm was retrained on a larger database". AI-Rad Companion (K213096), which analyzes lung CTs, references "additional training data was added", while Briefcase (K230020), a rib fracture triage device, mentions that
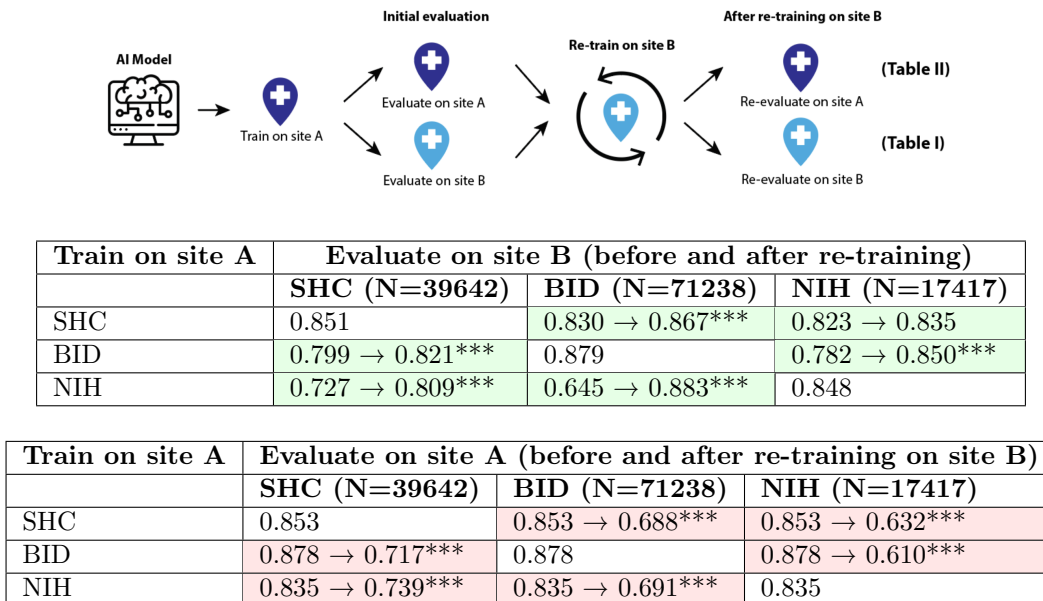
Train on site A	Evaluate on site B (before and after re-training)		
	SHC (N=39642)	BID (N=71238)	NIH (N=17417)
SHC	0.851	0.830 → 0.867***	0.823 → 0.835
BID	0.799 → 0.821***	0.879	0.782 → 0.850***
NIH	0.727 → 0.809***	0.645 → 0.883***	0.848

Train on site A	Evaluate on site A (before and after re-training on site B)		
	SHC (N=39642)	BID (N=71238)	NIH (N=17417)
SHC	0.853	0.853 → 0.688***	0.853 → 0.632***
BID	0.878 → 0.717***	0.878	0.878 → 0.610***
NIH	0.835 → 0.739***	0.835 → 0.691***	0.835

Figure 4: A case study on the effect of re-training pneumothorax AI models on other sites reveals that although re-training improves external site performance, performance consequentially degrades on the originally trained sites. **Top Figure:** An AI model is trained on pneumothorax cases from site A and then evaluated on held-out cases from site A and an external site (site B). Then, the model is updated by fine-tuning on 5K additional cases from site B and re-evaluated on sites A and B. This procedure is performed for three clinical sites (SHC, BID, NIH) across six total scenarios. The results of re-training and re-evaluation are shown in B and C. **Middle Table:** each cell shows the AUROC scores of the model evaluated on site B before and after re-training on site B. On average, models improved by 0.075 AUC after re-training. **Bottom Table:** each cell shows the AUROC scores of the model evaluated on site A before and after re-training on site B. Across both panels, we perform bootstrapped one-sided tests for each cell and indicate with asterisks (***) where $p < 0.001$.

the updated device differs "due to training the subject device on a larger data set". The remaining three devices reference the scale of the re-training dataset. For example, Quantib Prostate (K230772), which analyzes prostate MRIs, reports that the updated algorithm has been trained on "400 scans", while Genius AI (K221449), a breast cancer detection device, reports a "two-fold" increase. Finally, Caption Ejection Fraction (K210747), a cardiac ultrasound AI device, reports an "additional 30% training data from three ultrasound devices and two clinical sites". Details on these devices are also included in Figure 2. For the other 95 updated devices, we found several different update subtypes. The most common type of reported updating occurs when the manufacturer adds a new or additional prediction task to an existing model (55 total devices). For example, whereas FractureDetect's original device only works on wrists, its update has expanded to ankles, elbows, and other body parts. Next, we found that 21 devices have received updates to their accepted input signal. For example, recent mammography products such as Mammoscreen have included the ability to process Digital Breast Tomosynthesis (DBT)/3D scans, whereas previous versions only accepted Full-Field Digital Mammography (FFDM)/2D scans. An additional 13 devices report changes to the model design or architecture, such as a change from a fully connected neural network to a convolutional neural network. Five devices report a change to the intended target population for the device. For example, EndoSleep expanded its population to pediatric patients, whereas

the previous device only allowed for patients 18 or older. We found 22 devices that report changes to the model but do not specify the exact nature of the change. For example, approvals may report "additional algorithmic enhancements" or "improved quality of algorithms", but not reference whether the improvements come from re-training or model design. Finally, 37 devices report updates unrelated to the model or its usage. Namely, these include software or hardware changes that pertain to its interoperability or output interface. Examples include the UI/UX of the device which is visible to physicians, or a hardware configuration that allows the device to be installed on new machines. We provide a list of examples of these update types in Table 1.

3.2. Time Between Updates

Based on our dataset, updates of any type occur a median of 17 months after previous device approval, with follow-ups as short as 3.5 months and as long as six years (Figure 2). This is a relatively short window of time, as the median time from concept to FDA approval for non-AI medical devices has been estimated to be 31 months (C. Johnson et al., 2022). Additionally, in order to account for right-censorship in our dataset (e.g. not yet observed updates in the newer devices), we used the Kaplan-Meier estimator and produced its curve (Figure 2). At two years, devices have an estimated update probability of 20% for all update types, and at four years, this probability rises to 30%. After seven years, the estimated probability of update saturates at 35%, meaning that about a third of devices receive at least one update of any kind in their lifetimes. However, the reported rate of model re-training is significantly lower: within two years, 1.4% of models are reported to be retrained, with the probability of device updates saturating at 1.7% after 2.4 years.

3.3. Case Study

We carried out a case study to illustrate and quantify the tradeoffs with AI adaptation through model retraining. We investigated the potential benefits and challenges of re-training on additional data from external sites in pneumothorax AI algorithms (Wu et al., 2021b,a). We found that external evaluation of models can result in an AUC decrease of up to 0.18, while re-training and evaluating on data from external sites improves model performance in all scenarios, with an average of 0.075 and a maximum of 0.23 AUC

(Figure 4, Middle). However, after re-training on external sites, we also found that model performance degrades an average of 0.176 AUC (and up to 0.268 AUC) when re-evaluated on the original site (Figure 4, Bottom). This suggests that it can be challenging to have a single AI model that works well across heterogeneous settings.

4. Discussion

Currently, FDA-approved AI models are "locked" after approval, whereby making new changes requires undergoing a brand-new submission process, with most of the same regulatory burden (Gerke et al., 2020). Correspondingly, we observe in our analysis that only six out of 416 devices report actually received re-training updates, which is an essential approach for AI adaptation. On the other hand, nearly a quarter of devices receive updates in the form of additional marketing or functionality claims. Such disparity suggests a much stronger economic incentive for developers to increase the adoption of their devices through marketing new features rather than improving the original model through re-training.

One significant barrier to re-training is development costs, which may include acquiring new datasets (Chen et al., 2019; Wu et al., 2023a), computational resources (Wiens et al., 2019), data groundtruthing (Rahimi et al., 2021; Willemink et al., 2020), and regulatory hurdles (Kelly et al., 2019; Sertkaya et al., 2022). After models are updated, the manner in which they are deployed can also affect a device's ultimate clinical impact. First, while previous-generation AI devices for mammography were clinically evaluated to improve detection rates, subsequent studies showed limited benefits to women due to changes in how clinicians interacted with the devices, as well as the transition from film to digital mammograms (Lehman et al., 2015; Fenton, 2015). Second, economic forces such as reimbursement rates can affect how the frequency and extent to which these devices are adopted (Parikh and Helmchen, 2022b; Abràmoff et al., 2022). AI adoption is still in a nascent stage, with very few widely adopted products and underdeveloped commercial payment pathways (Chen et al., 2021; Parikh and Helmchen, 2022a; Wu et al., 2023b). In such an environment, companies with few customers may not be able to dedicate resources toward regular model updating

and maintenance. Currently, FDA cleared products exist in a similar band of risk profiles, with a previous study showing all devices currently categorized as risk class II (medium-risk) ((Zhu et al., 2022)). The lower-risk class I is largely exempt from the regulatory process, with the higher-risk class III reserved for devices that "sustain or support life, are implanted, or present potential unreasonable risk of illness or injury" ((for Devices and Health)). Whereas minimal-risk products like mobile health apps can introduce frequent updates without any regulatory hurdles, the medium-risk designation may encourage a trend towards more conservative updates that are more likely to be cleared rather than ambitious updates that may be rejected.

The FDA has recognized the high regulatory hurdles associated with model updating. In a recently proposed draft guidance from April 2023, model developers may be allowed to include a PCCP (Predetermined Change Control Plan) along with their device submission, which would allow them to simply document subsequent model updates rather than requiring a new submission every time (Center for Devices and Radiological Health, 2023b), potentially alleviating some of the regulatory burden and shortening update intervals. However, even under these proposed changes, developers are still required to complete rigorous evaluation and documentation of the algorithm changes, which incur much of the same prohibitive time and costs mentioned above (Allen, 2022; Evans, 2022). Furthermore, evaluating an updated model is an inherently difficult task due to various types of distribution shifts and heterogeneous data collection methods that are outside the control of developers (Schrouff et al., 2022; Chen et al., 2018). As such, future guidance documents should consider the challenges inherent in ensuring and evaluating fairness under fine-tuning and data shift. Our case study illustrates the tug-and-pull nature observed in AI models – when trained on data from a specific site, they can perform well, but this may trade-off with performance on other sites. Although our models are trained on only a few datasets and do not comprehensively represent the gamut of available training data sources and model architectures on the market, the results illustrate how one instantiation with commonly used data and architecture choices exhibits characteristic behaviors of performance shift. In the status quo, model developers are locked into one model, creating

scenarios where they may have to optimize for one population at the expense of another. To compound this issue, the actual performance on new, unseen populations is not even reported since the FDA does not require postmarket surveillance for 510(k)-approved devices (Wu et al., 2021a). To alleviate this problem, future regulatory guidelines should move beyond a "one-model-fits-all" approach, and instead consider allowing site-specific re-training and deployment. By allowing developers to deploy and validate multiple models under a single device, they can optimize model performance for each intended population without incurring performance tradeoffs. This would ensure that developers verify that their models perform well on each deployed clinical site while allowing them to perform the necessary site-specific documentation and evaluation as they mature. There are various design decisions that can affect how an AI model is re-trained: factors like whether to freeze layers, mix new training data, hyperparameter tuning, and validation processes can all influence how much re-training improves model performance (Pham et al., 2021; Picard, 2021; Qian et al., 2021). Indeed, in our case study, even though the individual models used in our ensemble approach only varied by the random seed used during training, performance across models still differed by up to 0.056 AUC. In a study by Watson et al. (2022), chest X-ray deep learning models trained on the same BIDMC dataset across different random seeds and hyperparameters were found to disagree in their explanations up to two-thirds of the time. Such studies on specific datasets represent potential pitfalls of algorithms applied to a particular clinical domain, but do not necessarily mean they generalize to all other types of devices. Regulatory guidelines should include consideration of appropriate fine-tuning schemes used when evaluating models.

Furthermore, we find that among models that have been updated with re-training, details on the data used in training are very limited, with basic descriptions such as "additional training data", or "larger database". A limitation of our study lies in the limited details reported in FDA clearances. For example, while only 6 devices report retraining on new data, 5 devices report updates to their target population and 21 devices report unspecified improvements to their algorithm. As such, the true rate of retraining on new data may be higher than reported. In order for consumers and users to make informed decisions on the impacts of model updates, regulators

should ensure that information about the data used for original model development and updates is transparent and accessible. Information such as patient demographics, hospital locations, disease subtypes, and healthcare settings are important covariates that can significantly influence model performance (Duffy et al., 2022; Wu et al., 2021b).

As the medical AI field matures, regulation should progress in lockstep with fully exploiting the technical benefits of adaptive learning systems while curbing risks to safety and efficacy. Looking beyond the US, regulatory bodies, such as the European Union, are similarly developing guidelines for regulating medical AI (Muehlematter et al., 2021). We believe that the trends and challenges in medical AI also extend to regulating other AI-transformed industries such as transportation and law, and offer important insights into how to appropriately foster AI innovation.

References

Michael D Abràmoff, Cybil Roehrenbeck, Sylvia Trujillo, Juli Goldstein, Anitra S Graves, Michael X Repka, and Ezequiel "Zeke" Silva III. A reimbursement framework for artificial intelligence in healthcare. *NPJ digital medicine*, 5(1):72, 2022.

Daphne Allen. AI, ML, & cybersecurity: Here's what FDA may soon be asking. https://www.designnews.com/artificial-intelligence/ai-ml-cybersecurity-heres-what-fda-may-soon-be-asking, April 2022. Accessed: 2023-7-15.

Boris Babic, Sara Gerke, Theodoros Evgeniou, and I Glenn Cohen. Algorithms on regulatory lockdown in medicine. *Science*, 366(6470):1202–1204, December 2019.

Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. 2018.

Larry J Brindza. What is a premarket notification 510(k)? *Clin. Microbiol. Newsl.*, 2(20):4–5, October 1980.

Center for Devices and Radiological Health. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices, October 2022. Accessed: 2023-8-16.

Center for Devices and Radiological Health. Product code classification database. https://www.fda.gov/medical-devices/classify-your-medical-device/product-code-classification-database, August 2023a. Accessed: 2023-8-30.

Center for Devices and Radiological Health. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial, March 2023b. Accessed: 2023-7-15.

Center for Devices and Radiological Health. Good machine learning practice for medical device development: Guiding principles. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles, October 2023c. Accessed: 2024-3-19.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

Melissa M Chen, Lauren Parks Golding, and Gregory N Nicola. Who will pay for AI? *Radiol Artif Intell*, 3(3):e210030, May 2021.

Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nat. Mater.*, 18(5):410–414, May 2019.

Gareth Corfield, Adam Mawardi, Hannah Boland, Daniel Woolfson, Melissa Lawford, and Eir Nolsøe. Tesla forced to update software in 1 million cars over 'sudden acceleration' fears. *The Daily Telegraph*, May 2023.

Grant Duffy, Shoa L Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng, and David Ouyang. Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit Med*, 5 (1):188, December 2022.

Nicholas Evans. A first look at the FDA's proposed regulatory framework for modifications to AI-based software as a medical device (SaMD): IP review and strategy guide. https://ipo.org/index.php/a-first-look-at-the-fdas-proposed-regulatory-framework-for-modifications-to-ai-based-software-as-a-medical-device-samd-ip-review-and-strategy-guide/, February 2022. Accessed: 2023-7-15.

Joshua J Fenton. Is it time to stop paying for computer-aided mammography? *JAMA internal medicine*, 175(11):1837–1838, 2015.

Center for Devices and Radiological Health. Learn if a medical device has been cleared by fda for marketing. URL https://www.fda.gov/medical-devices/consumers-medical-devices/learn-if-medical-device-has-been-cleared-fda-marketing#:~:text=43%25%20of%20medical%20devices%20fall,devices%20fall%20under%20this%20category.

Sara Gerke, Boris Babic, Theodoros Evgeniou, and I Glenn Cohen. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med*, 3:53, April 2020.

Stephen Gilbert, Matthew Fenech, Martin Hirsch, Shubhanan Upadhyay, Andrea Biasiucci, and Johannes Starlinger. Algorithm change protocols in the regulation of adaptive machine Learning-Based medical devices. *J. Med. Internet Res.*, 23(10): e30545, October 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 33 (01):590–597, July 2019.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III clinical database, 2023.

Kristin Johnson, Frank Pasquale, and Jennifer Chapman. Artificial intelligence, machine learning, and bias in finance: Toward responsible innovation symposium: Rise of the machines: Artificial intelligence, robotics, and the reprogramming of law. *Fordham Law Rev.*, 88(2):499–530, 2019.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.*, 17(1):195, October 2019.

David B Larson, Hugh Harvey, Daniel L Rubin, Neville Irani, Justin R Tse, and Curtis P Langlotz. Regulatory frameworks for development and evaluation of artificial Intelligence-Based diagnostic imaging algorithms: Summary and recommendations. *J. Am. Coll. Radiol.*, 18(3 Pt A):413–424, March 2021.

Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, Diana L Miglioretti, Breast Cancer Surveillance Consortium, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.

Urs J Muehlematter, Paola Daniore, and Kerstin N Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the USA and europe (2015-20): a comparative analysis. *Lancet Digit Health*, 3(3):e195–e203, March 2021.

Ravi B Parikh and Lorens A Helmchen. Paying for artificial intelligence in medicine. *NPJ Digit Med*, 5(1):63, May 2022a.

Ravi B Parikh and Lorens A Helmchen. Paying for artificial intelligence in medicine. *NPJ digital medicine*, 5(1):63, 2022b.

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: an analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, ASE '20, pages 771–783, New York, NY, USA, January 2021. Association for Computing Machinery.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. September 2021.

Shangshu Qian, Viet Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. Are my deep learning systems fair? an empirical study of fixed-seed training. *Adv. Neural Inf. Process. Syst.*, 34: 30211–30227, 2021.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. March 2019.

Saba Rahimi, Ozan Oktay, Javier Alvarez-Valle, and Sujeeth Bharadwaj. Addressing the exorbitant cost of labeling medical images with active learning. In *International Conference on Machine Learning in Medical Imaging and Analysis*, page 1, 2021.

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine Heller, Silvia Chiappa, and Alexander D'Amour. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. February 2022.

Aylin Sertkaya, Rebecca DeVries, Amber Jessup, and Trinidad Beleche. Estimated cost of developing a therapeutic complex medical device in the US. *JAMA Netw Open*, 5(9):e2231609, September 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew B A McDermott, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest x-ray classifiers. In *Pacific Symposium on Biocomputing 2021*, pages 232–243. unknown, November 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 875–884. IEEE, January 2022.

W Bradley Wendel. Technological solutions to human error and how they can kill you: Understanding the boeing 737 max products liability litigation. *J. Air L. & Com.*, 84(3):379–446, 2019.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.*, 25(9):1337–1340, September 2019.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295 (1):4–15, April 2020.

Bernd W Wirtz, Jan C Weyerer, and Benjamin J Sturm. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, 43(9):818–829, July 2020.

Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.*, 181(8):1065–1070, August 2021.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.*, 27(4):582–584, April 2021a.

Eric Wu, Kevin Wu, and James Zou. Explaining medical AI performance disparities across sites with confounder shapley value analysis. November 2021b.

Kevin Wu, Dominik Dahlem, Christopher Hane, Eran Halperin, and James Zou. Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations. In Bobak J Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C Ho, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 229–242. PMLR, 2023a.

Kevin Wu, Eric Wu, Brandon Theodorou, Weixin Liang, Christina Mack, Lucas Glass, Jimeng Sun, and James Zou. Characterizing the clinical adoption of medical AI through U.S. insurance claims. August 2023b.

Simeng Zhu, Marissa Gilbert, Indrin Chetty, and Farzan Siddiqui. The 2021 landscape of fda-approved artificial intelligence/machine learning-enabled medical devices: An analysis of the characteristics and intended use. *International journal of medical informatics*, 165:104828, 2022.