Locked Artificial Intelligence? Why Regulators Should Use an Ongoing Monitoring Approach for Machine Learning-Based Medical Devices

Authors: Boris Babic[1], Sara Gerke[2], Theodoros Evgeniou[3], I. Glenn Cohen[4]*†

Affiliations:

[1]Boris Babic, INSEAD, France and Singapore; California Institute of Technology (Caltech), Pasadena, California 91125, USA.

[2]Sara Gerke, The Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School; The Project on Precision Medicine, Artificial Intelligence, and the Law (PMAIL), Cambridge, MA 02138, USA.

[3]Theodoros Evgeniou, INSEAD, France and Singapore.

[4]I. Glenn Cohen, Harvard Law School, Cambridge, MA 02138, USA.

†All authors contributed equally to the analysis and drafting of the paper.

*Correspondence to: igcohen@law.harvard.edu.

Abstract: Regulators of medical Artificial Intelligence and Machine Learning (AI/ML) are faced with a difficult problem: Should they limit marketing to a version of the system that was submitted for initial premarket review (a "locked" regime), or permit marketing of an algorithm that can adapt to changing conditions (an "adaptive" regime)? In April 2019 FDA issued a draft framework to address this problem that may become a model worldwide. In this Policy Forum, we argue that the locked/adaptive distinction and FDA's proposed approach to it miss the central risks of medical AI/ML. Such risks emerge from structural features of predictive analytics, like concept drift, covariate shift, and AI/ML model instability. Paying attention to these risks suggests a continuous and cooperative monitoring framework of how the systems work in different and likely evolving environments, which we outline by way of conclusion.

One Sentence Summary: Regulators should adopt an ongoing risk management process for medical machine learning updates rather than focus on the locked/adaptive distinction.

Medical Artificial Intelligence and Machine Learning (AI/ML) is a fast-growing area with many technologies already hitting the market: from mobile apps that help visually impaired to better interact with their environment (*1*) or detect skin cancer (*2*), to an AI/ML tool that analyzes chest X-rays to identify suspected findings suggestive of pneumothorax (*3*), to clinical use of IDx-DR, the first AI/ML diagnostic that provides a screening decision for the eye disease diabetic retinopathy (*4*). Medical AI/ML is big business, with a recent forecast suggesting that the market for these technologies will surpass $34 billion worldwide by 2025 (*5*). As regulators like the United States (U.S.) Food and Drug Administration (FDA), the national competent authorities of the European Member States, and the European Medicines Agency struggle with how to regulate medical AI/ML, they face a very fundamental problem: after reviewing a submission, should the regulator limit its marketing authorization to a version of the algorithm (and underlying training data) that was submitted (what FDA calls a "locked" algorithm, a terminology we employ in this paper while recognizing it may not be ideal) or permit marketing of an algorithm that can learn and adapt to new conditions (what FDA calls an "adaptive" algorithm)? We refer to this as the AI/ML Update Problem. For drugs and ordinary medical devices, this problem typically does not arise. But AI/ML is unique in this regard, as it is the first such technology with the capability to continuously evolve.

In this Policy Forum, we address the Update Problem by analyzing and critiquing the approach suggested by FDA in its April 2019 proposed framework (*6*). While crediting the value of FDA's approach, we also explain why the problem is more nuanced and difficult than FDA's proposal suggests and show how a framework sensitive to risks that are unique to and ubiquitous in

modern ML may prove superior. Our message, in short, is that regulators should focus on whether the AI/ML system as a whole is appropriately stable, with emphasis in particular on treating similar patients similarly.

The Regulatory Design Problem

One of the key advantages of AI/ML is that it can enable a "learning healthcare system," wherein the boundaries between research and practice are regarded as porous (*7*). Once the AI/ML is deployed, it can (to anthropomorphize slightly) learn and thereby alter its performance/behavior, much the way a medical resident learns on the job. But this poses a difficult regulatory design challenge; to see it, it is useful to begin with two polar approaches to the Update Problem.

The first pole would be for a regulator to permit marketing of only a "locked" algorithm and require any change to the algorithm to undergo a completely new premarket review. Such an approach has several drawbacks. Suppose an algorithm for analyzing the results of mammograms and making recommendations as to breast cancer risk receives marketing authorization (*8*). Suppose the training data was under-inclusive of African-American women who tend to have differences in breast density from Caucasian women. The algorithm would thus produce recommendations ill-suited for that population. As the AI/ML system is used in clinical settings that include more African-American women, it becomes possible to more accurately estimate the parameters used to predict breast cancer in this sub-population when making recommendations. Moreover, in some situations, AI/ML identifies sub-populations that were *not* known ex-ante. For example, in conducting HIV vaccine studies, researchers did not know (and

perhaps could not know) ex-ante that in a particular vaccine trial the vaccine might increase rather than reduce HIV infection risk for "uncircumcised men who both had sex with men (MSM) *and* had high levels of pre-existing Ad5 antibodies" (*9*). Indeed, with usage, the AI/ML can even develop customized models for different sub-populations (some of which are only possible to identify after using it on lots of patients) as it accumulates data about them -- AI/ML is exciting in part because it sometimes draws relationships that would not be anticipated by human beings. Such customization would be health-promoting, but if another premarket review is needed, the update may never occur -- the maker may not have a financial incentive to pursue the cost of another review and might also worry about what message pursuing it might send about the quality of its existing algorithm.

The opposite pole would be to treat the initial marketing authorization as permitting the AI/ML maker to update the algorithm without any further regulatory review. Such updates can be either of the algorithm itself ('algorithm updates') -- for example, replacing a linear ML model with a polynomial one -- or of the algorithm's parameters ('parametric updates') which may be continuously tuned as the system is applied to new data in practice. This approach is likewise perilous. Parametric updates are at the core of modern AI/ML systems -- they take place almost continuously, without human input, and their effects can be hard to identify ex-ante. But the quality of parametric updates depends largely on the quality of the associated underlying data. An adaptive system with continuously changing parameters is susceptible to data quality issues that can arise from, for example, errors of the AI/ML users or adversarial attacks on machine learning -- as described by Finlayson *et al.* in a recent *Science* Policy Forum (*10*). The latter can

take many forms. Consider a hypothetical example, as in (*10*): in response to the opioid crisis, many insurance companies now use patient or provider level overdose risk prediction algorithms to deny oxycontin prescription filings. A physician, certain that she has a patient in need of a prescription, may learn that she can avoid the algorithmic gatekeeper and secure a prescription by typing in a combination of codes which will guarantee a low-risk for overdose score. Such a system incentivizes the elicitation of low-quality physician data. An unchecked dynamic algorithm would inappropriately adapt to this and begin to falsely categorize low-risk patients as high risk. The algorithm's evolution in this context is analogous to the way that Tay, Microsoft's AI chatbot, learned to post inflammatory and racist tweets in response to adversarial attacks from Twitter users feeding it 'low-quality data' in the form of incendiary speech (*11*). In this kind of situation, ongoing oversight of the sort we will encourage can provide a necessary check on adaptive AI/ML systems.

FDA's Proposed Framework

In an attempt to steer between the Scylla and Charybdis of these two polls, FDA released a discussion paper in April 2019 (*6*). Until now, FDA has only approved or cleared medical AI/ML devices with "locked" algorithms (*6*). A "locked" algorithm is defined by FDA as "an algorithm that provides the same result each time the same input is applied to it and does not change with use." (*6*). Any AI/ML system can satisfy this definition provided it is fixed in advance.

However, most AI/ML algorithms are "adaptive", arguably their key strength. For example, even parameters in a simple model like a logistic regression will gradually evolve as we refit the model in response to new data. For adaptive AI/ML-based software as a medical device – what FDA calls "SaMD", which is software that is on its own a medical device and is *not* part of a hardware medical device (*12)* – FDA proposed in its discussion paper a "total product lifecycle (TPLC) regulatory approach" that permits the continuous improvement of such devices while maintaining their safety and effectiveness (*6*). The idea is that AI/ML systems could be updated to a certain extent after marketing authorization; when seeking initial premarket review of an AI/ML-based SaMD, manufacturers would be given the option to submit a "predetermined change control plan", which would contain a description of anticipated modifications and an "Algorithm Change Protocol", including the associated methodology being utilized to implement such changes (*6*).

Understanding the Risks of Medical AI/ML

In its discussion paper of April 2019, FDA adds to its existing SAMD approach the idea of a spectrum between locked and adaptive algorithms. We argue that this distinction can be misleading, and even dangerous, by cultivating a false sense of control: an algorithm which the FDA defines as "locked" could be more harmful than an "adaptive" one – and vice versa. Instead, for approving AI/ML-based devices and their subsequent updates, the FDA should focus on whether the AI/ML system as a whole is overall reliable. We will describe below several types of AI/ML risks that can undermine the system's reliability.

To begin, the concept of "locked" is not well defined in FDA's approach. Consider, by analogy, a targeting system following an aircraft. To "lock", in this context, may be defined for the system to focus on a fixed point in space, rather than on its moving target. Doing so, when the system fires at times 1 and 2 based on the image of the fixed point (*identical inputs*), it will strike the exact same location (*identical outputs*). However, the target airplane itself has moved. Further, its location may be still uncertain due to imperfections of the radar itself. Under these conditions, locking (to geographic coordinates) is not helpful – indeed, it is counter-productive. Such a system would not be effective. What the targeting system must do, instead, is to *track the moving target*. Extending the analogy to medicine, what we want is for the AI/ML system to lock, as closely as possible, to the true function that relates the inputs and outputs – which is unknown in practice – rather than continuing to use the estimate of that function that was first approved (see Box 1). Locked systems might, quite literally, miss the mark. We identify below several types of risks that, when properly controlled, can help the AI/ML system lock on to the true function as closely as possible. These risks include (1) *concept drift*, (2) *covariate shift*, and (3) *model instability*. This is where we believe the FDA should focus its attention.

Box 1: Classification and Risk

The goal in a typical machine learning task is to predict an unknown label $Y \in A$ from a number of features or covariates $\mathbf{X} \in V$, where A denotes the outcome space and V the feature space. This is a function estimation problem: $Y = f(\mathbf{X})$. In practice, the true function f is not known. The goal is to identify an estimate of that function that best predicts the true value of Y from the data **X**, where best is defined in terms of minimizing the error rate or, more generally, a loss function: ℓ: $V \times A \rightarrow \mathbf{R}$.

Consider $h = \Pr(Y|\mathbf{X})$. While we do not ordinarily know this important quantity, the theoretically best classifier would be $g(\mathbf{X}) = \text{argmax}_x\, h(\mathbf{X})$ that assigns each observation to the class for which h is largest. We may also prefer to receive the output in terms of h alone. In either case, concept drift refers to a situation where h changes over time:

$$h^{(t1)}(\mathbf{X}) \neq h^{(t2)}(\mathbf{X}) \quad \textbf{(Concept Drift)}$$

Covariate shift refers to a situation where the distribution of **X** alone changes

$$\Pr^{(t1)}(\mathbf{X}) \neq \Pr^{(t2)}(\mathbf{X}) \textbf{ (Covariate Shift)}$$

Following the approach of Dwork *et al.*'s seminal paper on fairness (*13*), for any two patients, consider d: $V \times V \rightarrow \mathbf{R}$, a measure of divergence between any two individuals; M: $V \rightarrow \Delta(A)$, a mapping from individuals to probability distributions over outcomes; and $D : Q \times Q \rightarrow \mathbf{R}$, a measure of divergence between two probability distributions. To treat similar patients similarly, we would require that for every pair of patients $\mathbf{x}, \mathbf{y} \in V$

$$D(Mx, My) \leq d(x, y) \textbf{ (Lipschitz Property)}$$

An AI/ML system that does not satisfy the Lipschitz property is *not stable*.

Concept Drift

In AI/ML, concept drift describes a situation where the relation between inputs and outputs changes over time. This may happen due to a changing environment or because the model was mis-specified (e.g., the estimated function is linear when the actual relationship is quadratic).

Consider, for example, an AI/ML system trained to identify skin lesions as benign or malignant, as in Esteva *et al.* (*14*). The model presupposes an underlying distribution of these labels (benign vs. malignant). However, the datasets these AI/ML systems rely on, such as the ISIC dermoscopic archive (https://isic-archive.com/), typically do not track race or skin color, even if tests were done with all skin colors. Yet the malignancy of skin lesions may vary across race and skin type. As a result, the same image can lead to two different probabilistic diagnoses, depending on the underlying skin/race, an omitted feature. This problem is ubiquitous in medical AI/ML. Locking the algorithm does not protect against such harms, much like locking the coordinates in the aircraft example above does little to ensure that the target is hit. Indeed, a locked algorithm can make matters worse by prohibiting the system from learning from experience.

Meanwhile, while FDA's discussion paper on adaptive AI/ML-based SaMD is a welcome development, its proposed predetermined change control plan is either uninformative or impractical – depending on the level of detail at which a maker would be expected to describe future modifications. At one extreme, we might require a maker to describe proposed changes in very general (and hence impractical) terms. This would be uninformative. On the other extreme, we might require them to describe precisely the sorts of changes they anticipate. Even if such a task can be accomplished, such a plan could be harmful when we learn about unanticipated problems – in which case, the proposed framework could require another round of review. Thus, such an approach would be impractical.

Covariate Shift

When the input distribution (distribution of new data) is different from the data the algorithm was trained or tested for approval on, we have covariate shift (*15*). This can occur in the absence of concept drift. For example, it may be that our training data came from geographically centralized clinical sites. When this occurs, locking the algorithm hampers the maker's ability to address the problem. Further, describing how the marginal distribution may change is not something a maker may be able to do *ex-ante* since they usually do not know the distribution of the data that the algorithm will be applied to.

Instability

As discussed above, one major concern is treating similar patients similarly. Suppose that when an AI/ML system is given a set of inputs, it produces one probabilistic output. For example, the probability that a particular skin lesion is malignant is 87%. Now suppose that we make very small changes to the set of inputs provided to the underlying algorithm. For example, if the input space is a high dimensional pixel space, where the algorithm is given an image of the lesion, we may change the values associated with a few pixels in a way that is medically insignificant. The AI/ML system must now make a prediction from a feature vector that is vanishingly different from $\boldsymbol{x}$: $\boldsymbol{x} + \boldsymbol{\delta}x$. For most reasonable metrics d, the distance between $\boldsymbol{x}$ and $\boldsymbol{x} + \boldsymbol{\delta}x$ is close to 0.

A stable algorithm should give predictions that are similarly "close" in the output space (in probability) when it is given $\boldsymbol{x} + \boldsymbol{\delta}x$ instead of $\boldsymbol{x}$ alone. In other words, if $d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta}x) \cong 0$ then $D(\mathrm{M}x, \mathrm{M}\boldsymbol{x} + \boldsymbol{\delta}x) \cong 0$. More generally, we would require that $D(\mathrm{M}x, \mathrm{M}\boldsymbol{x} + \boldsymbol{\delta}x) \leq d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta}x)$

(Dwork *et al.* (*13*) use this property as the basis for a definition of fairness in ML, see Box 1). When this inequality is not satisfied, the algorithm is not stable in the sense that similar patients can receive dissimilar diagnoses. From the perspective of patient safety, we would not want a diagnostic system that frequently classifies medically similar lesions very differently. Paying attention to the Lipschitz Property (see Box 1) encourages us to think not in terms of same inputs/same outputs, but in terms of similar inputs/similar outputs. Encouraging this shift is among our central insights. As Ralph Waldo Emerson put it, *foolish consistency is the hobgoblin of little minds.*

In modern AI/ML, leading classifiers are highly nonlinear. This makes them especially vulnerable to such instability, as demonstrated, for example, in Goodfellow *et al.* (*16*), where $\boldsymbol{\delta}x$ is taken to be an imperceptibly small vector. In (*10*), Finlayson *et al.* focus on how such instabilities can lead to adversarial attacks in medical applications. Indeed, they demonstrate that popular skin cancer classification algorithms are often unstable. But the problem extends beyond adversarial attacks. Locking an algorithm does not secure against instability and FDA's proposed framework, while moving in the right direction, does not get to the core of the problem either, because it is impossible to know in advance what *kind* of instabilities the world actually has.

An Ongoing Monitoring Approach for Medical AI/ML Systems

For the reasons indicated above, we argue that FDA's current approach (locked/adaptive distinction) is not well suited to the problem. We believe FDA should instead focus on the central risks of AI/ML, as identified above. While this is in the spirit of FDA's proposed TPLC

approach (as we explain below), the emphasis should be on developing a *process* to identify and manage risks rather than on articulating a plan for updating *ex-ante*. Such a process can include, for example, the following aspects:

- **Retesting:** An AI/ML system may need to be regularly retested on all past cases, including the ones used for the initial FDA market authorization. Major discrepancies on past verdicts may lead to regulatory action.
- **Simulated checks:** An AI/ML system should be continuously applied to 'simulated patients' in order to evaluate whether its behavior is reliable with respect to a sufficient diversity of patient types.
- **Adversarial stress tests:** Every AI/ML system may need to be paired with an adversarial monitoring mechanism (*17*). A stable AI/ML system should be robust to the kinds of adversarial modifications described in Finlayson *et al.* (*10*). The FDA should therefore use the adversarial approach to conduct algorithmic stress tests throughout the AI/ML system's lifecycle.
- **An appropriate division of labor:** Monitoring of AI/ML systems should, in general, be done by actors different from the ones developing these systems. Separation of development and testing is common in other contexts: for example, in software development, quality assurance and development teams are separate, while risk management and compliance departments are separated from traders in the financial sector. Such divisions may be likewise required for companies developing medical AI/ML systems. Moreover, third-party organizations that monitor AI/ML systems based

on standards the industry develops, similar in spirit to those of professional organizations like the IEEE and the ISO, may also play a role in the future.

- **Use of electronic systems:** Finally, regulators could use electronic systems to continuously monitor AI/ML systems using testing procedures as outlined above. For example, the FDA's national medical product monitoring system *Sentinel* (*18, 19*), mainly used for identifying risks from usage of drugs, vaccines, and other biologics, could be enhanced to continuously monitor the behavior of approved AI/ML devices. Combining information from Electronic Health Records (EHRs) and other data from such devices, the FDA could itself perform some of the tasks described above in a manner patterned on *Sentinel*.

The above suggestions potentially complement FDA's proposed new TPLC approach, which we commend and seek to build on. More generally, however, our goal is to move away from a framework of identifying anticipated changes and emphasize the risks that can arise from unanticipated changes in how medical AI/ML systems adapt to their environments. As described above, subtle often unrecognized parametric updates can cause large and costly mistakes. The focus for such updates should not be on identifying or documenting them, which may be challenging if not impossible, but instead on managing risks that are specific to this environment (concept drift, covariate shift, and model instability). Finally, while our discussion has focused on FDA and the U.S. experience, with appropriate adaptations, the ongoing monitoring approach we set out can be potentially used by other countries and their regulators as well.

Box 2: Concluding Insights

Regulators **should not** focus on the locked/adaptive distinction. Rather, they should adopt an appropriate risk management framework, focusing in particular on whether similar inputs lead to similar outputs.

Regulators **should not** focus on the submission of a plan or change control protocol *ex-ante*. It is usually not possible to know what we don't know.

Regulators **should** focus on establishing a monitoring and risk management regime for medical AI/ML systems. Such a regime would include:

Retesting,

Simulated checks,

Adversarial stress tests,

An appropriate division of labor, and

Use of electronic systems.

References and Notes:

1. OrCam (https://www.orcam.com/en/).
2. G. A. Zakhem, C. C Motosko, R. S. Ho, How should artificial intelligence screen for skin cancer and deliver diagnostic predictions to patients? *JAMA Dermatol.* **154**, 1383-1384 (2018).
3. U.S. Food and Drug Administration, (FDA), "Zebra Medical Vision Ltd." (K190362, Health PNX, 2019; https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190362.pdf).
4. M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, J. C. Folk, Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine* **1**, 1-39 (2018).
5. Tractica, "Artificial intelligence for healthcare applications" (2019; https://www.tractica.com/research/artificial-intelligence-for-healthcare-applications/).
6. U.S. Food and Drug Administration, (FDA), "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, 2019; https://www.fda.gov/media/122535/download).
7. The learning healthcare project, "Background, learning healthcare system" (2019; http://www.learninghealthcareproject.org/section/background/learning-healthcare-system).

8. A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60-66 (2019).
9. A Fauci, M. Marovich, C. Dieffenbach, E. Hunter, S. Buchbinder, Immune activation with HIV vaccines: implications of the adenovirus vector experience. *Science* **344**, 49-51 (2014).
10. S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, I. S. Kohane, Adversarial attacks on medical machine learning. *Science* **363**, 1287-1289 (2019).
11. E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter" (2016, https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter).
12. For more information on the term "SaMD", see U.S. Food and Drug Administration, (FDA), "Software as a medical device (SaMD)" (2018; https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd).
13. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS), Cambridge, MA, 2012.
14. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).
15. S. Bickel, M. Bruckner, T. Scheffer, Discriminative learning for differing training and test distributions, Proceedings of the 24th International Conference on Machine Learning (ICML), Corvallis, OR, 2007.

16. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, 2015.
17. S. Gu, L. Rigazio, Towards deep neural network architectures robust to adversarial examples, Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, 2015.
18. U.S. Food and Drug Administration, (FDA), "FDA's Sentinel Initiative" (2019; https://www.fda.gov/safety/fdas-sentinel-initiative).
19. Sentinel Coordinating Center, "Sentinel is a national medical product monitoring system" (2019; https://www.sentinelinitiative.org/).

Acknowledgments:

Funding:

S.G. and I.G.C. were supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784).

Competing interests:

I.G.C. served as a bioethics consultant for Otsuka on their Abilify MyCite product. The authors declare no other competing interests.