Comment on Ariel Dora Stern, "Regulation of Medical AI: Policy Approaches, Data, and Innovation Incentives"

Boris Babic

December 2022

1 Introduction

In the development of medical artificial intelligence (medical AI) applications there exists a pressing set of open questions around how to effectively build the associated regulatory landscape. Among these questions is what the role of the US Food and Drug Administration (FDA) and its counterparts in other countries should be, and how much we should now rely on the institutional infrastructure that historically evolved for the oversight of traditional (non AI) software in medical devices. In "Regulation of Medical AI: Policy Approaches, Data, and Innovation Incentives", Stern develops a very illuminating project that can help us begin to chart answers to these questions. The project is backed by an insightful empirical analysis – using FDA data on cleared medical AI devices together with data on adverse events and recalls – that sheds further light on who is developing these technologies, how they are performing, and their relative safety.

In this comment I will offer (1) a partial summary of Stern's article, (2) a brief analysis of the interesting empirical insights, and (3) a more open-textured discussion of the emerging trends and unsettled questions in medical AI regulation.

2 Partial Summary

While there is a large variety of medical AI applications, as Stern explains, substantial energy and attention centers on those that are performing or assisting in diagnostic

and treatment tasks. These are likely to qualify as medical devices under the 1976 Medical Device Amendment to the 1938 Federal Food, Drug, and Cosmetic Act. The FDCA is the chief statute empowering the FDA with authority to oversee the safety of medical products.

When a technology qualifies as a medical device, the FDA takes a tri-partite approach to its regulation. The lowest risk (Class I) devices are subject only to modest (manufacturing) controls. Moderate risk (Class II) devices are regulated through a process called premarket notification (or the 510(k) pathway), requiring the maker to demonstrate either substantial equivalence to an existing regulated device or to pass a de novo classification request. The de novo process in turn requires the maker to provide reasonable assurance of safety and effectiveness of the device for its intended use. Finally, the highest risk (Class III) devices require premarket approval, which typically includes evidence from clinical trials.

When it comes to medical AI, most devices are dubbed Software as a Medical Device by the FDA (SaMD) – as opposed to software in a medical device (SiMD) and are going through the moderate risk (Class II) processes. For example, Benjamens et al. (2020) compile a database of medical AI technologies that have been cleared by the FDA, and of the 79 devices on their list all but two were brought to market under the Category II scheme – i.e., going through the 510(k) or de novo pathway.

Accordingly, Stern's empirical analysis runs through the following pipeline: Stern begins with the full FDA 510(k) database for the years 2010-2022Q3, containing over 38,000 devices. The scope is then further limited to applications in eight of the largest medical specialties, resulting in approximately 31,000 clearances. This database is then merged with two additional sources of information, (1) data from the FDA's medical device adverse event reporting database, and (2) data from the FDA's recall database. Stern then uses text analysis to identify devices with a software component ($\sim$8,500 devices), and among those Stern uses additional keywords to identify AI based software devices as a proper subset (303 devices). Importantly, the keywords used to identify AI based devices are "artificial intelligence," "deep learning," "machine learning," and "neural network."

3 Empirical Insights

Some interesting trends worth highlighting are the following: the use of the Class II de novo pathway is twice as common among AI devices; the majority of devices are being brought to market by privately held firms; and the number of clearances has

more than doubled per year during the observation period. Geographically, the United States, Israel and Japan hold a disproportionate share of the medical AI innovation market. While it is small surprise that the United States has the most FDA clearances, the outsize performance of Israel and Japan relative to its peer countries with strong biomedical innovation, such as France and South Korea, is interesting.

Perhaps the most notable, however, is Stern's analysis of device safety outcomes – made possible by the merging of 510(k) clearance data with data on adverse event reports and recalls. While it is very insightful it is also necessarily preliminary – because we simply do not have enough devices, or a long enough observation period, to make more definitive statements about the relative safety of medical AI vs. non-AI technologies.

For example, in the full sample of approximately 30,000 devices, there were slightly over 4,000 adverse event reports. For the subset of devices containing software (AI or non AI) there were just over 1,000 adverse event reports, out of a total of approximately 8,500 devices. And for AI based medical devices, there were 9 adverse event reports out of a total of approximately 300 devices. Meanwhile, for recalls, there are 1,000 recalls in the full sample, just over 500 in the subset of software based devices, and 5 in the subset of AI based software devices. While this suggests that, proportionately, medical AI devices are overall safer, we must temper that conclusion by how much we can learn from the small numbers observed over a relatively short period. As an aside, assessments of adverse event reports also have a censoring problem worth taking into consideration – we cannot distinguish between the non-occurrence of an event and an event that occurred but was unreported.

There is also an interesting methodological question here. Consider a hypothetical example: In medical materials engineering, emerging technologies are often quite invasive – such as catheters, VADs, and heart stents. They are used to treat very serious illness, and their role in the body can be critical to a patient's survival. As a result, the potential for things to go wrong is substantial. And when things do go wrong, they go wrong unambiguously – for example, a device breaks in the patient's body or stops pumping blood as it should. This is not an ideal example, because such devices would likely be Class III devices, but I use it here merely to illustrate a general point – namely, that in non medical AI devices used for treatment, it is clear how defects can occur and it is likewise clear what would constitute evidence of such defects.

But now consider some archetypal medical AI applications: For example, consider an imaging diagnostic assistant tool – a device that, say, reads an x-ray and outputs a probability of a bone fracture. What would be a mistake or a malfunction in this

case? After all, the result is produced in the form of a probability. And, it is used in conjunction with a radiologist's expert opinion. So unless the device crashes, it is hard to envision a situation where we would see evidence of a defect from its performance.

One thing we might do is try to stress test the device in an adversarial fashion – try to identify cases where small changes in input lead to large changes in output, as Babic et al. (2019) suggest. This would be closer to a defect, because the classification function is failing to satisfy a Lipschitz condition, so to speak – cases that are similar along some metric in their inputs are treated very differently in their outputs. But this is not something that users would do in the ordinary application of medical AI devices. And that further suggests that maybe in the world of medical AI, we will need to focus more on ongoing regulation and assessment, as Gerke et al. (2020) argue, than on traditional adverse event reports or user identified problems leading to recalls.

4 Emerging Trends and Open Questions

This leads to the final section. I will focus on three open questions: (1) updating, (2) model transparency, and (3) regulatory loopholes.

4.1 Updating

The above considerations suggest that traditional approaches to evaluating safety performance may not be ideal as applied to medical AI devices. This is consistent with what Babic et al. (2019) dub the update problem. Traditionally, the FDA has required software based medical devices to undergo a new round of review every time the underlying code is changed.

As Stern explains, for a Class II device, there are no regulatory provisions for amending or changing an existing 510(k) clearance, and any modification would presumably require a new 510(k) to be submitted. Class III devices require a "PMA supplement", an onerous submission justifying the software changes. In other words, once SaMD is approved, the associated software is locked on approval.

This makes for a very unproductive regulatory approach for medical AI, where the main benefit comes from the algorithm's ability to learn from new data. For instance, imagine a simple linear classification function where the odds of x are given by $e^{\beta_0+\boldsymbol{\beta}x'}$. As part of the approval process the β parameter coefficients would be fit to some training sample. Now, as the algorithm is applied in practice and new observations come in, would a change in the βs trigger a requirement for a new 510(k)? Plausibly yes, because

any change in the βs can change the input-output relationship, which the FDA requires to be fixed. Such a policy is very antithetical to the 'learning' in 'machine learning.'

Fortunately, as Stern recognizes, the FDA has recently proposed a new total product life-cycle regulatory approach, which would move away from the black and white practice of approve/deny and its associated discouragement of software updates. Since this proposal is still in its infancy, it is hard to know how it will look, but in principle it could make for a more productive partnership between the FDA and medical AI manufacturers. Indeed, it could allow regulators to move away from looking at isolated adverse event reports and to take a more participatory and ongoing monitoring role of medical AI devices – to take a system view, as Gerke et al. (2020) suggest. For example, regulators can be on the lookout for common modeling problems that can lead to patient harm, such as concept drift, covariate shift, and model instability (in the sense of similar inputs leading to very different outputs) (Babic et al., 2019).

4.2 Model Transparency

Stern identifies AI devices using the keywords enumerated above – such as "deep learning" and "neural network". These keywords are typically associated with so-called "black box" machine learning models, and it is worth considering whether there are other medical AI applications that do not use these terms which are missed by the search – for example, devices described as classification models, multivariate analyses, regressions, or statistical learning techniques. I doubt there are many, but if there are, it would be particularly illuminating to include them because such models are more likely to be transparent ("white boxes") and a number of scholars have argued that they should be preferred in medicine (Babic et al., 2021a) and other high stakes settings (Rudin, 2019). Indeed, it would be interesting to compare the performance of different types of medical AI devices (black box vs. white box) with respect to adverse event reports and recalls. It may be that white box models have less adverse event reports, but it may also be that problems are easier to identify when white box models are used, with the counter intuitive implication that they have mode adverse event reports without being any less safe.

This brings up a more general question of how if at all model transparency should enter into the regulatory equation. Currently, the FDA is agnostic between different types of classification algorithms. That is, there is no necessary advantage to using a transparent linear model as opposed to a deep learning one, from the perspective of gaining FDA clearance for a medical AI product. But as the agency transitions to the total product life-cycle regulatory approach, it is worth considering whether the black

box nature of a system's algorithm is something to be on the lookout for.

4.3 Loopholes

While significant attention has been paid to Class II and III medical AI devices, very few medical AI applications actually qualify as a device, and even if they might qualify as a device, the FDA often exercises enforcement discretion, meaning that the FDA will not enforce regulatory requirements over these products. In effect, then, they are altogether outside the scope of the FDA's regulatory purview. This is true in particular for what are deemed "health or wellness" applications (Babic et al., 2021b). For example, mobile apps that are designed to track weight and fitness levels.

Devices like these pose an interesting problem from a public policy perspective. Arguably we tolerate the manufacturers circumventing the regulatory landscape because the individual risk they pose to patients is low. For example, consider mobile phone apps that have a partial diagnostic function – such as detecting high heart rate. Since these apps tend to have high sensitivity the biggest risk, one might think, is that of a false positive (a false health scare) requiring a specialist follow-up.

However, it is worth asking whether our regulatory infrastructure should be built around individual patient harm in the way that the regulation of medical practice is. As policy makers, we may want to look at the medical AI ecosystem as a whole. And from a social/aggregate perspective, these negligible individual costs can add up. For example, if millions of people require a specialist follow-up to correct a false positive generated by a mobile phone app, this creates a large social cost borne by taxpayers. Another way to put the point: when it comes to unregulated mobile apps, perhaps device manufacturers should be required to bear the costs of the health care externalities that they generate.

5 Concluding Remarks

At the widest level of generality, and by way of closing, it is worth asking whether we should regulate algorithms by their domain of application (medicine, criminal justice, finance, etc.), or whether we should have one agency that regulates algorithmic technologies across different domains, as Tutt (2017) has recently argued. Stern effectively demonstrates how the former approach, which is currently the one we take, requires agencies to significantly upgrade regulatory environments that were developed a long time ago and for very different purposes. And we see this struggle in the case of the FDA – attempting to quickly evolve their approach to governing medical software in a way that can suitably cover medical AI applications. Arguably, the latter approach

(a separate agency for governing algorithms across domains) would allow for a more uniform, flexible and holistic regulatory environment for all AI technologies, regardless of their field of application.

References

Babic, B., S. Gerke, T. Evgeniou, and I. Cohen (2019). Algorithms on Regulatory Lockdown in Medicine. *Science 366*(6470), 1202–1204.

Babic, B., S. Gerke, T. Evgeniou, and I. Cohen (2021a). Beware Explanations from AI in Health Care. *Science 373*(6552), 284–286.

Babic, B., S. Gerke, T. Evgeniou, and I. Cohen (2021b). Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications. *Nature Machine Intelligence 3*, 283–287.

Benjamens, S., P. Dhunnoo, and M. Bertalan (2020). The State of Artificial Intelligence-Based FDA- Approved Medical Devices and Algorithms: An Online Database. *NPJ Digital Medicine 3*(1), 1–8.

Gerke, S., B. Babic, T. Evgeniou, and G. Cohen (2020). The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device. *Nature Digital Medicine 3*(53).

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence 1*, 206–215.

Tutt, A. (2017). An FDA For Algorithms. *Administrative Law Review 69*(1), 83–123.