Dr. AI, Where Did You Get Your Degree?

Edward Raff(✉), Shannon Lantzy(✉), and Ezekiel J. Maier(✉)

Booz Allen Hamilton, McLean, USA
{raff_edward,lantzy_shannon,maier_ezekiel}@bah.com

Abstract. Federal health agencies are currently developing regulatory strategies for Artificial Intelligence based medical products. Regulatory regimes need to account for the new risks and benefits that come with modern AI, including safety concerns and unique opportunities, like the potential for autonomous learning, that makes AI dramatically different from traditional static medical products. The current default regulatory regime is to treat AI like a medical device (i.e., as opposed to like a drug or a biologic product). As agencies like the U.S. Food and Drug Administration (FDA) develop new regulation to cover the uniqueness of AI, we suggest they consider adopting aspects of regulation traditionally used in the practice of medicine (i.e., doctors). In fact, FDA is currently undergoing a pilot that moves in that direction. We propose that AI regulation in the medical domain can analogously adopt aspects of the models used to regulate *medical providers*. We provide this view point to encourage discussion of how medical AI might be regulated. In doing so, we will also review several issues our framework does not resolve.

Keywords: Regulation · Continuous learning · Clinical applications

1 Introduction

Governmental agencies like the FDA are anticipating a wave of new software products for medical applications, and are currently drafting regulatory guidance in anticipation of this wave. Goals of new regulatory guidance include protecting the public from risk, reducing the time to market for these devices, and fostering an innovative market for the new software. For example, the FDA's Digital Health Program is running an nine-company pilot program[1] to pre-certify organizations developing software as a medical device (SaMD) for streamlined premarket review [1]. However, FDA's recent draft publication[2] stops short of providing guidance for artificial intelligence as a medical device (AIaMD). In this

[1] https://www.fda.gov/medicaldevices/digitalhealth/digitalhealthprecertprogram/default.htm.

[2] https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM524904.pdf.

The views expressed in this paper come solely from the authors, and do not represent an official or endorsed position by Booz Allen Hamilton.

F. Koch et al. (Eds.): AIH 2018, LNAI 11326, pp. 76–83, 2019.
https://doi.org/10.1007/978-3-030-12738-1_6

paper, we extend the SaMD discussion to a regulatory framework for AIaMD. For the purpose of this paper, we define AI-enabled medical device as a software product that actively learns after it is released to the market, and that is intended to inform or make decisions on behalf of a health care provider or patient.

Medical products, such as drugs, biologics, and non-AI devices undergo an evidence-based review of their safety and efficacy, i.e., their benefit-risk profiles. AIaMD will upend this traditional regulatory paradigm because, by definition, the devices can automatically change their own benefit-risk profile without human intervention. For example, an algorithm to detect cancer from MRI imagines and recommend treatment pathway could become more precise and sensitive over time by learning from cases in situ. While we have not yet seen reports of AIaMD in the health care market, it is crucial that governments provide clear guidance on how upcoming AIaMD product submissions will be reviewed and approved. Promising AI-enabled medical products have surfaced, albeit ones that do not continuously learn. For example, in early 2017, Arterys Inc. received FDA 510(k) clearance for its web-based medical imaging analytics software[3]. The lack of AIaMD submissions may be due to lack of sufficient readiness of the technology, but it may also be stymied by the lack of clear regulatory guidance and government approval pathways. The development of clear AIaMD regulation will provide market stability and encourage innovation due to: (1) improved consumer confidence in the safety and efficacy of products; (2) a clear understanding of the requirements for marketing approval, thereby allowing companies to judge risk of their investment going to market, and informing academic and institutional review boards of the requirements surrounding medical studies. As AI researchers, it is critical we have a voice in how this regulation forms to ground expectations and ensure that innovation is not unduly stifled.

We believe there is a risk that harmful regulation could be established (i.e., regulation that does not increase safety and efficacy but prevents or slows innovation) due to fear and the uncertainty around AIaMD. For example, the often "black-box" nature of AI has spurred considerable demand for interpretability and explainability in an AI-based medical device [10]. A "right to explanation" has already been codified in the European Union's laws [7]. Regulatory review of medical products traditionally focuses on evidence of *safety* and *effectiveness* over interpretability or mechanism of action. We contend that *mandating* interpretability is excessively burdensome for AI-enabled devices. This is not to say that interpretability has no value; AI systems that can explain their choices may warrant faster regulator approval. But to focus on interpretability as a necessity for AI would stifle progress.

Rather than focusing regulation on algorithm explainability and self-updating models, we would like to shift focus to outcomes for the patient and to the healthcare market. In this paper we use the paradigm of regulating the practice of medicine as a framework for thinking differently. We propose elements of a framework analogous to the standards used to license medical providers.

[3] https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare.

Similar to accredited *medical schools* which train medical doctors, we contend that AI-enabled devices should be trained utilizing accredited *data collection and validation methods*. AIaMD trained using these accredited data collection and validation methods should then be evaluated based on measured outcomes to individual patients. Similar to state medical boards which remove harmful doctors from practice, we contend that we need an AI regulator to surveil and remove AIaMD if they become harmful.

2 Regulatory Design for AIaMD

To ensure that the immense potential of AI is not hampered, stakeholders must actively engage in the development of the regulatory framework. Researchers, software product developers, patient advocates, medical providers, and payers' participation in this discussion will help to avoid the hype and fear that has led to previous AI winters. We argue that the methodological accreditation and outcomes-focus framework outlined below, will enable regulatory agencies to accomplish their mandate of protecting public health while allowing for innovation by AI researchers. However, discussion, dialogue, and iteration is needed. The FDA has invited public feedback and participation in the conversation.[4]

2.1 Accrediting Our Data Sources and Methods

Doctors are educated by accredited universities. AIaMD should be trained with accredited data and methods. While much of the discussion around AI focuses on the algorithms used, data collection and the training methods are extremely important to the success of any model. AI is not immune to the "garbage-in garbage-out" problem, and so ensuring that high-quality algorithms are developed means we must ensure data is of an equally high quality. Accrediting the process by which data is acquired and prepared provides the foundation needed for any level of trust in the results. Accreditation of a dataset's labeling and creation process should mirror the acceptance criteria of sufficient evidence for new clinical guidance in medical practice. For example, the dataset accreditation scheme should consider: an appropriate diversity of patient backgrounds (e.g., age, BMI, etc); a diversity of feature sources (e.g., MRI images used for training must come from multiple MRI machines of differing versions and differing vendors); the consistency of feature sources between the training and clinical contexts; the completeness of data meta-information; defined measurable and clinically-relevant outcomes (e.g., real-time insulin levels), rather than measures that may be available (e.g., unqualified claims records). Fully satisfying all of these goals may not be possible in each case, but should always be considered and addressed. Significant failures in any of these sub-components can prevent development of actionable and effective AI solutions. For example, [12] found

[4] https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm\#getinvolved.

that out of 2,511 recent genome-wide association studies, 81% of all participants were of European ancestry. This poses a risk that developed solutions and results will be ineffective for the majority of the world's population.

As part of Booz Allen Hamilton's organization of the 2016 and 2017 Data Science Bowl competitions [2], which focused on detecting heart function and lung cancer respectively, organizers examined each of these aspects of the competition data to ensure that it was high-quality and enabled the development of useful algorithms. We found unexpected metadata which artificially boosted the algorithm's appearance of clinical performance (i.e., leakage). Specifically, meta-information describing the hospital that labeled the cardiac MRI images proved to be strongly predictive of a specific heart measurement, despite having no clinical diagnostic power. If this meta-information was not recorded, organizers would not have discovered the correlated, but not actionable feature, and could have led to model overfitting to the training data. This exemplifies why data should be acquired from a diversity of locations, and why trained medical providers must be part of the data preparation process. As one step toward ensuring the safety, AI-enabled devices must be robust to a diversity of input sources. The best way to achieve this robustness is to utilize a diverse high-quality data set for training.

It is possible for regulators to take a proactive approach by creating gold-standard data sets for important and prevalent conditions. Such data could be used in multiple ways to both improve the efficiency of regulation and the speed at which products are developed. These could be kept as secret evaluation sets to confirm reported performance, an independent training set to independently test system generalizability, or even provided to product developers to reduce data acquisition costs and promote marketplace competition. The FDA is already exploring the development and curation of a standard dataset for radiogenomics [8]. This could also allow the FDA to preemptively remove barriers that slowed the adoption of Electronic Medical Records in the United States relative to other nations, such as lack of capital and standardized data exchange formats [3].

2.2 Focus on the Outcomes

Doctors' outcomes are monitored by their medical boards, colleagues, and patients; AIaMD postmarket surveillance should include a diversity of feedback sources. By definition, AIaMD learn from well-defined outcomes which are measured while in use. Therefore, post-market surveillance (i.e., monitoring the benefit-risk profile of a medical product after it has been released on the market) can be built directly into an AI product. AIaMD developers should focus on building a system capable of collecting the right outcomes. Regulators should focus on the process by which an AI device developer defines, collects, and uses post-market outcomes to refine and improve the model. Next, similar to a doctor who is subject to review and possible sanctions by their state medical board (i.e., probation periods with added surveillance, or suspension from medical practice), regulators should sanction and/or withdraw an AIaMD from the market for egregious errors.

We propose that, like medical review boards for medical providers, regulators should institute AI review boards consisting of a multidisciplinary group of experts from within and outside the regulatory agency. The AI boards would include continuing education-like requirements to update AI models using new standards and ground truth data, sanctioning AI producers for errors or AI misconduct or bias, and removal of an AI product when it does harm. Trials and studies will remain necessary to ensure that the device is both safe (does no harm), and effective (provides meaningful and quantifiable improvement in outcomes).

3 So Can We Treat AI Like a Doctor?

Framing the regulation of AI in the same manner as medical doctors provides a basis for constructing regulation for non-static products. This approach allows regulators, the AI community, and the general public to debate the opportunities and obstacles of AI-enabled medical devices.

A primary psychological benefit of this approach is to avoid the problem of moving goal posts or an AI double standard. The public is often unwilling to trust a machine to perform a task unless the outcome is *far better* than what a human can produce.[5] This thought process ignores the intrinsic benefits of availability and faster decision making. For example, AI-enabled medical devices can provide both routine care in rural and poor communities that would have no access otherwise, and faster diagnosis, leading to improved patient outcomes. With regulation focused on data accreditation and clinical outcomes, regulators avoid unnecessarily delaying adoption of AI technology for medicine.

This regulatory framework also provides guidance on ensuring AI devices remain safe over time. Physicians are not simply told to do no harm. Rather, physicians progress from interns to specialist over their careers, and as they progress their responsibilities and autonomy increases. AI devices could follow a similar (task-dependent) progression. This lends to a natural encouragement for AI products to be developed in an incremental approach. However, AI devices need not progress completely to autonomous continually learning agents (i.e., a specialist). Instead AI devices can ultimately be tools, which have utility to physicians irrespective of their autonomous continually learning capability.

With this regulatory approach we must collectively recognize that errors and mistakes will be made. Just as doctors, drugs, and devices sometimes unintentionally harm patient, AIaMD will as well. Just as deaths due to medical errors occur, so do deaths caused by software bugs [9]. Every death is tragic; yet the question of safety is not whether a doctor or an AIaMD prevents all harm, but rather he/she/it reduces the rate of harm from the current standard of care. SaMD deaths in Leveson [9] were incidents that the FDA studied in order to remediate and prevent future incidents. While the hope for AI devices is to reduce the frequency of such unfortunate incidents, the same lessons will apply to the AI space. Researchers who acquire and prepare the data, and develop

[5] https://phys.org/news/2016-05-humans-automated-advisor-bad-advice.html#jCp.

models to analyze take action must understand this risk. Given the potential greater autonomy of AIaMD, AI developers may require a form of "malpractice" insurance. This insurance would provide fiscal and regulatory incentives to encourage safety and provide financial recompense when incidents occur.

4 Failure Points

We believe lifting and adapting from the regulatory framework for medical providers is useful to frame our discussion around regulating AI. However, that framework is not perfect as it exists today, and we see no reason to expect it will be perfect for AI either. It is important to also discuss points where the regulatory schema for the practice of medicine will not work for AIaMD. In the sections below we discuss these failure points and offer prompts to develop thought and discussion from the community. Below we will discuss three issues, which we feel are important toward developing complete regulation.

4.1 Recalling AI

The reach of bad AIaMD will be broader than the reach of a bad doctor. Every year thousands of doctors are sanctioned by their state medical boards. Morrison and Wickersham [11] found that 79% of California's disciplinary cases resulted in some form of license suspension or revocation. This is an important issue and part of the reason physicians are licensed, but it is also reactive—action does not occur until something goes wrong. During the time between misconduct and revocation, these doctors are unfortunately putting their patients at risk. Similarly, some AIaMD products will need to be recalled (i.e., have their "license" suspended) in the same reactive manner. We will again have an issue with the time between product failure ("misconduct") and a successful removal from the market. But in this case, an AI product could have potentially been deployed nation wide or even globally, where a single doctor's misconduct is intrinsically limited to a smaller pool of people. This increases the potential cost (e.g., of patient well being, potential monetary damages) of an AI failure case.

AIaMD may be less fungible than individual doctors, making removal more disruptive. Removing AIaMD may also be more locally disruptive than removing a bad doctor because it may be too unique. If one doctor is removed from medical practice, there are other doctors who can step in to perform the functions. However, if AIaMD performs a unique function that becomes an essential part of a clinical workflow, it may be more difficult to replace the function. For example, if radiologists begin to rely heavily on computer analysis of tumor images, removing that AIaMD may cause a temporary lapse in care for tumor analysis.

4.2 Adversarial AI and Security

While the medical industry has long had to handle sensitive personally identifiable and protected health information, security of this information has not

historically been reviewed by regulatory agencies. The Health Insurance Portability and Accountability Act (HIPAA) laws in the U.S. provide some regulation regarding security issues, and will require updates as AIaMDs enters the market. With the advent of adversarial machine learning, a new kind of security issue must also be considered.

In adversarial machine learning, a hypothetical adversary attempts to trick a classifier into making specific incorrect decisions [5]. This research field, which is at the intersection of machine learning and computer security, is most prevalent in the fields of spam filtering, malware detection, and computer vision community, including for self-driving cars. Due to the potential to interact with adversaries, AI-enabled medical device developers must also consider this form of attack. Notably, AI-enabled medical device adversarial interactions may be with individuals engaging in drug-seeking behaviour, as well as sophisticated malicious groups. Fraud is already an enormous issue in the medical field, with hundreds of billions of dollars lost, and there is fear that this problem will only worsen with the adoption of machine learning systems [6].

Ultimately, it is not yet known to what degree adversarial attacks will affect SaMD and AIaMD. AIaMD developers can follow current practices of defining a threat-model by which adversaries can act to evaluate the risk to their systems [4]. However, it has so far been found that such attacks are easy to create and apply, even with threat-models that are highly restrictive to the adversary's actions and knowledge [5]. Regulators must eventually decide how far AIaMD developers must go to protect systems from attacks, and determine in advance domains where their product should not be applied due to risk of attack. This is an issue that will require careful consideration, and by its very nature, not one that we can rely on current systems to handle.

5 Conclusion

Fundamentally, medical regulation exists precisely because without it consumers cannot reasonably assess the quality of all possible medical diagnoses and the benefits and risks of recommended treatments. Regulatory agencies are developing new policy and guidance for static SaMD, and will soon codify rules to govern dynamic AIaMD. Rather than developing new regulations based on our existing rules for static medical products, we proposed using the analogy of medical practice regulation as a foundation to develop a novel regulatory framework for AI-enabled devices. We argue that the regulatory framework for medical practice provides a natural paradigm to address the public's concerns about the use of AI in healthcare, and we have used it to illustrate points of consideration for new regulation. Though the accreditation process for medical doctors is not perfect, the approach has served society for decades and can serve as the foundation for regulating AI-enabled medical devices.

Acknowledgments. We thank anonymous reviewers as well as participants in AIH 2018 workshop at FAIM for their generous feedback and discussions, which have improved this paper.

References

1. Digital health innovation action plan introduction. Technical report, Food & Drug Administration (2017). https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf
2. Data Science Bowl (2018). https://datasciencebowl.com/
3. Anderson, J.G.: Social, ethical and legal barriers to e-health. Int. J. Med. Inform. **76**(5), 480–483 (2007). https://doi.org/10.1016/j.ijmedinf.2006.09.016. http://www.sciencedirect.com/science/article/pii/S1386505606002218. ISSN 1386-5056
4. Biggio, B., Fumera, G., Roli, F.: Security evaluation of pattern classifiers under attack. IEEE Trans. Knowl. Data Eng. **26**(4), 984–996 (2014). https://doi.org/10.1109/TKDE.2013.57. ISSN 10414347
5. Biggio, B., Roli, F.: Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning, pp. 32–37 (2017). http://arxiv.org/abs/1712.03141
6. Finlayson, S.G., Kohane, I.S., Beam, A.L.: Adversarial Attacks Against Medical Deep Learning Systems (2018). https://arxiv.org/abs/1804.05296
7. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Mag. **38**(3), 50 (2017). https://doi.org/10.1609/aimag.v38i3.2741. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2741. ISSN 0738-4602
8. Gottlieb, S.: FDA's comprehensive effort to advance new innovations: initiatives to modernize for innovation. Technical report, Food and Drug Administration (2018). https://blogs.fda.gov/fdavoice/index.php/2018/08/fdas-comprehensive-effort-to-advance-new-innovations-initiatives-to-modernize-for-innovation/
9. Leveson, N., Turner, C.: An investigation of the Therac-25 accidents. Computer **26**(7), 18–41 (1993). https://doi.org/10.1109/MC.1993.274940. http://ieeexplore.ieee.org/document/274940/. ISSN 0018-9162
10. Lipton, Z.C.: The doctor just won't accept that! In: Interpretable ML Symposium at NIPS (2017). http://arxiv.org/abs/1711.08037
11. Morrison, J., Wickersham, P.: Physicians disciplined by a state medical board. JAMA **279**(23), 1889 (1998). https://doi.org/10.1001/jama.279.23.1889. http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.279.23.1889. ISSN 0098-7484
12. Popejoy, A.B., Fullerton, S.M.: Genomics is failing on diversity. Nature **538**(7624), 161–164 (2016). https://doi.org/10.1038/538161a. http://www.nature.com/doifinder/10.1038/538161a. ISSN 0028-0836