Regulatory Oversight, Causal Inference, and Safe and Effective Health Care Machine Learning

Biostatistics (2020) **21**, 2, *pp.* 363–367
doi:10.1093/biostatistics/kxz044
Advance Access publication on November 19, 2019

Regulatory oversight, causal inference, and safe and effective health care machine learning

ARIEL DORA STERN*

Harvard Business School and the Harvard-MIT Center for Regulatory Science, Morgan Hall 433, 15 Harvard Way, Boston, MA 02163, USA

astern@hbs.edu

W. NICHOLSON PRICE II

University of Michigan Law School, 625 State Street, Ann Arbor, MI, USA and the University of Copenhagen Center for Advanced Studies in Biomedical Innovation Law (CeBIL), Copenhagen, Denmark

Summary

In recent years, the applications of Machine Learning (ML) in the health care delivery setting have grown to become both abundant and compelling. Regulators have taken notice of these developments and the U.S. Food and Drug Administration (FDA) has been engaging actively in thinking about how best to facilitate safe and effective use. Although the scope of its oversight for software-driven products is limited, if FDA takes the lead in promoting and facilitating appropriate applications of causal inference as a part of ML development, that leadership is likely to have implications well beyond regulated products.

Keywords: Causal inference; Device regulation; FDA; Machine learning; Software as a medical device.

In recent years, the applications of machine learning (ML) in the health care delivery setting have grown to become both abundant and compelling. Tools with applications ranging from the diagnosis of diabetic retinopathy (FDA, 2018a; Gulshan *and others*, 2016) to software that generates an alert for urgent pneumothorax findings in chest X-rays regularly make headlines, while multiple areas of medicine ranging from psychiatry to cardiology have seen the debut and use of new ML-driven tools with potential to improve patient care and clinician decision-making (Bzdok and Meyer-Lindenberg, 2018; Cohen, 2019; Johnson *and others*, 2018).

In the United States, regulators have taken notice of these developments and the U.S. Food and Drug Administration (FDA) has been engaging actively in thinking about how best to facilitate safe and effective use. The FDA regulates to ensure the safety and efficacy of many health care products, including medical devices that incorporate software and the subset of those devices with ML capabilities.

Although it may not be immediately apparent that ML-driven tools fall under the FDA's purview, the definition of a "medical device" is broad and includes any "instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article which is intended for use in the

*To whom correspondence should be addressed.

diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals" (21 U.S. Code § 321). The FDA's definition thus includes not only physical devices with ML software components (or software resulting from ML development) embedded in the devices—known as software *in* a medical device, or SiMD—but also stand-alone software itself, known as software *as* a medical device, or SaMD (International Medical Device Regulators Forum, 2014). Such software-driven products present unique challenges—for example, cybersecurity and the management of data collection and privacy—as well as unique opportunities for improving patient care (Gordon and Stern, 2019).

Notably, FDA does not actively regulate two significant types of ML systems: certain types of clinical decision support software (CDS) and laboratory-developed tests. CDS helps providers make care decisions by, for instance, providing suggested drug dosages or alerts about drug interactions. Under § 3060 of the 2016 21st Century Cures Act, certain CDS that gives providers the opportunity to review the rationale behind its recommendations is exempt from the definition of a medical device; FDA typically cannot regulate such software (21 U.S. Code § 360j). Additionally, some ML systems are "laboratory-developed tests" that are developed and used within a single health care setting, such as Hopkin's TREWS and Duke's SepsisWatch system for monitoring and predicting sepsis (Henry *and others*, 2015; Futoma *and others*, 2017). While FDA has the authority to regulate such tests, it has long held back from exercising its authority over them (FDA, 2018b).

FDA regulators evaluating SaMD consider manufacturer-claimed functionality when assessing its risk profile in the regulatory review process. Under an international framework developed with the assistance of FDA regulators, SaMD is grouped into risk categories based on two features: how serious the health condition is (non-serious, serious, or critical), and how significant the SaMD's output is to the health care decision (i.e., whether it informs clinical management, drives clinical management, or treats or diagnoses directly) (FDA, 2017). FDA has emphasized the role of clinical and analytical validation in ensuring that SaMD performs safely and effectively.

FDA has also thought proactively about the special considerations raised by ML in SaMD products. In April 2019, FDA released a discussion paper on a "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device." Some critics had raised early concerns that FDA would follow rigid premarket approval procedures that would slow the ability of ML products to improve with the availability of new data and new models (Price II, 2017). However, under the 2019 proposal, FDA put forth a framework that would allow ML product developers to establish procedures by which ML products can be updated more flexibly and with fewer regulatory hurdles, including with the use of real-world data and real-world evidence (FDA, 2019a). Further, FDA officials have spoken publicly about the importance of a "reimagined approach" to regulating algorithms (Patel, 2019).

Many substantial challenges remain in the development of safe and effective ML products. The generalizability of ML identifications, predictions, and recommendations remains questionable, particularly in situations where ML fails to demonstrate causality—or is simply not designed to do so, as is the case for many diagnostic tools, where algorithm developers only asking the ML to discover predictive rather than causal relationships. Subbaswamy and Saria (2020) highlight aspects of these generalizability challenges further, discuss the statistical foundations for why these concerns arise, and present some potential remedies. Data curation in individual health care settings is frequently *ad hoc* and context specific, which raises concerns for applying algorithms outside these settings (National Academy of Medicine, 2017). Transfer learning—defined as "the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned" (Torrey and Shavlik, 2010)—also remains difficult (Wiens *and others*, 2014). Biases present in training datasets can affect the resulting algorithms, arising from bias in the provision of health care itself or non-representativeness of the training environment in terms of patient populations or care-provision resources (Robinson *and others*, 2020; Price II, 2019;

Williams and Wyatt, 2015). Causal estimation techniques can help ameliorate some of these challenges by mitigating false attribution, though such techniques cannot fully address them.

In this complex environment, the role of regulatory leadership in ensuring safe and effective ML in health care tools is particularly important. One of the key ways in which regulators support the development of new regulated medical products, such as drugs, is through the qualification of biomarkers (FDA, 2019b). According to the Biomarkers, EndpointS and other Tools (BEST) glossary, compiled by the FDA and National Institutes of Health, a biomarker is “a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions” (FDA-NIH Biomarker Working Group, 2016). Even in the absence of formal regulatory qualification (which has traditionally been the domain of drug, rather than device regulation), biomarkers require validation, which has been defined as the process of conducting “formal assessments that show that measuring the biomarker is a reliable indicator of a patient’s clinical status or outcome” (Hey *and others*, 2019).

For software-driven products, researchers have begun to rely on *digital biomarkers*. These are measures that are especially compelling in that they “can support continuous measurements outside the physical confines of the clinical environment” (Coravos *and others*, 2019). Many digital biomarkers will be familiar to patients and clinicians, as they represent digital and/or remotely collectable versions of biomarkers that have historically been collected in the absence of software, such as continuous blood glucose meters with data transmission capabilities. Other digital biomarkers represent algorithms that quantify entirely novel outcomes—for example, voice analysis software that can predict the progression of neurological diseases and smartphone sensing of mobility to predict depression.

While biomarker validation is important everywhere in the clinical research endeavor, and while regulators have historically played a key role in formal biomarker qualification, *causal inference* may be especially important in the establishment and validation of novel digital biomarkers. For example, an ML algorithm that can predict which patients are likely to benefit from a therapy may not be relevant in settings that differ meaningfully from those in which the algorithm was trained *unless* causal inference tools were used in development (Shalit, 2020). In some cases, this will require more algorithmic transparency than certain ML techniques afford, as researchers push to understand the factors that drive ML algorithms to make specific diagnoses or treatment recommendations.

Notably, the development of digital biomarkers shares many of the more general challenges of biomarker development with respect to incentivizing parties in the health care system to innovate and participate (Stern *and others*, 2018). For example, although they may benefit from validated biomarkers, health care providers are largely compensated for—and therefore focused on—the delivery of health care services rather than R&D activities. And like other exercises in biomarker validation, the validation of novel ML algorithms can have long-lasting and far-reaching implications for future product development. As such, the thoughtful application of tools for causal inference in ML should be a priority for both the developers of novel digital biomarkers and the regulars who may rely on such biomarkers in evaluating the safety and effectiveness of future products. Fortunately, as other articles in this collection highlight (Rose and Rizopolous, 2020), the statistician’s toolkit has been growing rapidly and several rigorous techniques have been developed for causal inference in ML in the biomedical setting.

If FDA takes the lead in promoting and facilitating appropriate applications of causal inference as a part of ML development, that leadership is likely to have implications well beyond the scope of regulated products. For example, although many tools that apply ML in health may not go through an FDA regulatory approval or clearance process, including laboratory-developed tests and rationale-providing clinical decision support software as described above, these products must still be safe and effective in order to create value for patients and clinicians. Therefore, efforts by regulators to validate digital biomarkers and other ML algorithms using the full statistical toolbox of causal inference techniques (Diaz, 2020) may help further the safety and efficacy of non-regulated products by establishing standards in a nascent industry. Thus, the thoughtful development of ML-based medical devices may have positive spillovers that encourage quality and establish causal inference techniques in health care ML more broadly.

Digital biomarkers have the potential to accelerate efforts to create, validate, and improve ML products in health care, whether marketed medical devices or software developed and used solely in-house by health care delivery organizations. Recent discussions suggest that FDA leaders are engaging thoughtfully with open questions about the applications of ML to regulated products as well as discussing the importance of standardized language and the establishment of "good machine learning practices" (Joseph, 2019). All stakeholders, from FDA, to developers, to health systems, should encourage careful use of causal inference tools in developing those digital biomarkers to ensure that ML algorithms are safe and effective.

Acknowledgments

The authors are grateful to Lila Kelso for editorial assistance. A.D.S. gratefully acknowledges support from the Kauffman Junior Faculty Fellowship, and W.N.P. gratefully acknowledges support from the Novo Nordisk Foundation (NNF17SA0027784).

Conflict of Interest: None declared.

References

Bzdok, D. and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **3**, 223–230.

CDER Biomarker Qualification Program. (2019). U.S. Food and Drug Administration website: http://www.fda.gov/drugs/drug-development-tool-qualification-programs/cder-biomarker-qualification-program (August 2, 2019).

Cohen, T. (2019). *Israel's Zebra Medical gets FDA ok for AI chest X-ray product. Reuters.* https://www.reuters.com/article/us-healthcare-zebra-medical-regulator-idUSKCN1SJ0LI.

Coravos, A., Khozin, S. and Mandl, K. D. (2019). Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *Npj Digital Medicine*, **2**, 1–5.

Diaz, I. (2020). Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* **21**, 353–358.

FDA-NIH Biomarker Working Group. (2016). *BEST (Biomarkers, EndpointS, and other Tools) Resource.* http://www.ncbi.nlm.nih.gov/books/NBK326791/.

Futoma, J., Hariharan, S., Sendak, M., Brajer, N., Clement, M., Bedoya, A., O'Brien, C., and Heller K. (2017). An improved multi-output Gaussian process rnn with real-time validation for early sepsis detection. In: *Proceedings of Machine Learning for Healthcare 2017*. arXiv:1708.05894.

Gordon, W. J. and Stern, A. D. (2019). Challenges and opportunities in software-driven medical devices. *Nature Biomedical Engineering* **3**, 493–497.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A. *and others* (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410.

Henry, K. E., Hager, D. N., Pronovost, P. J. and Saria, S. (2015). A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, **7**, 299ra122.

Hey, S. P., D'Andrea, E., Jung, E. H., Tessema, F., Luo, J., Gyawali, B. and Kesselheim, A. S. (2019). Challenges and opportunities for biomarker validation. *Journal of Law, Medicine & Ethics* **47**, 357–361.

International Medical Device Regulators Forum. (2014). Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations. p. 30.

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M. *and others* (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology* **71**, 2668–2679.

Joseph, A. (2019). *Q&A: The FDA's digital health chief on how to regulate futuristic AI products. STAT News.* https://www.statnews.com/2019/08/09/fda-artificial-intelligence-regulation/.

National Academy of Medicine. (2017). *Optimizing Strategies for Clinical Decision Support.* https://nam.edu/optimizing-strategies-clinical-decision-support/.

Patel, B. (2019). *Software Precertification: A Reimagined Approach.* https://drive.google.com/file/d/1N5gV1jpLj10UK7lXhcxCQ9wlsL4T7nLp/view?usp=embed_facebook.

Price II, W. N. (2017). Regulating black-box medicine. *Michigan Law Review* **116**, 421.

Price II, W. N. (2019). Medical AI and contextual bias. *Harvard Journal of Law & Technology.* **33**, https://papers.ssrn.com/?abstract_id=3347890.

Robinson, W. R., Renson, A. and Naimi, A. I. (2020). Teaching yourself about structural racism will improve your machine learning. *Biostatistics* **21**, 339–344.

Rose, S. and Rizopolous, D. (2020). Machine learning for causal inference in *Biostatistics*. *Biostatistics* **21**, 336–338.

Shalit, U. (2020). Can we learn individual-level treatment policies from clinical data? *Biostatistics* **21**, 359–362.

Stern, A. D., Alexander, B. M. and Chandra, A. (2018). Innovation incentives and biomarkers. *Clinical Pharmacology & Therapeutics* **103**, 34–36.

Subbaswamy, A. and Saria, S. (2020). From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA: IGI Global, pp. 242–264.

U.S. Food and Drug Administration (2017). Software as a Medical Device (SAMD). Clinical Evaluation - Guidance for Industry and Food and Drug Administration Staff. https://www.fda.gov/media/100714/download.

U.S. Food and Drug Administration. (2018a). FDA Permits Marketing of Artificial Intelligence-based Device to Detect Certain Diabetes-related Eye Problems. U.S. Food and Drug Administration website: http://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye (August 12, 2019).

U.S. Food and Drug Administration. (2018b). Laboratory Developed Tests. U.S. Food and Drug Administration website: http://www.fda.gov/medical-devices/vitro-diagnostics/laboratory-developed-tests (August 2, 2019).

U.S. Food and Drug Administration. (2019a). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback*. https://www.fda.gov/media/122535/download.

U.S. Food and Drug Administration. (2019b). About Biomarkers and Qualification. https://www.fda.gov/drugs/cder-biomarker-qualification-program/about-biomarkers-and-qualification.

Wiens, J., Guttag, J. and Horvitz, E. (2014). A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association: JAMIA* **21**, 699–706.

Williams, D. R. and Wyatt, R. (2015). Racial bias in health care and health: challenges and opportunities. *JAMA* **314**, 555–556.

[*Received September 25, 2019; revised September 25, 2019; accepted for publication September 25, 2019*]