Trust, but Verify: Informational Challenges Surrounding AI-Enabled Clinical Decision Software

The CENTER for INNOVATION POLICY at DUKE LAW

Duke
MARGOLIS CENTER for Health Policy

Authors

Christina Silcox, Managing Associate, Duke-Margolis Center for Health Policy

Arti Rai, Elvin R. Latty Professor of Law and Co-Director, The Center for Innovation Policy, Duke University School of Law

Isha Sharma, Senior Research Assistant, Duke-Margolis Center for Health Policy

Acknowledgements

The authors would like to deeply thank many people for contributing their time and expertise to inform and improve this white paper. The paper would not have been possible without the input of expert perspectives from a May 2019 private workshop and January 2020 public meeting, both held in Washington, DC. Additionally, the Duke team held many informational meetings and calls with various stakeholders across the artificial intelligence and health care ecosystem over the course of this project. We are extremely grateful for their time and thoughtful feedback on working drafts.

We would also like to thank Patricia Green, the Duke-Margolis Center's Director of Communications, for her editorial support and Kerry Stenke from the Duke Clinical Research Institute for her support in developing graphics for this white paper. Finally, we would like to thank Duke University's Balfour Smith, Bennett Wright, Michael McCarthy, and Nikhil Gadiraju for their contributions to the research and editing of this report.

Funding

This publication was funded by the Greenwall Foundation.

Disclosures

Any opinions expressed in this paper are solely of those of the authors and do not represent the views or policies of other organizations external to Duke.

About the Duke-Margolis Center for Health Policy

The Robert J. Margolis, MD, Center for Health Policy at Duke University is directed by Mark McClellan, MD, PhD, and brings together expertise from Washington, DC, the broader policy community, Duke University, and Duke Health to address the most pressing issues in health policy. The mission of the Duke-Margolis Center is to improve health and the value of health care through practical, innovative, and evidence-based policy solutions. Duke-Margolis catalyzes Duke University's leading capabilities, including interdisciplinary academic research and capacity for education and engagement to inform policy-making and implementation for better health and health care. For more information, visit healthpolicy.duke.edu.

About the Center for Innovation Policy at Duke Law

The Center for Innovation Policy at Duke Law (CIP), led by Faculty Co-Directors Arti Rai and Stuart Benjamin, is a forum for independent analysis of policies for promoting technological innovation that enhances long-term social welfare. CIP brings together technology and business leaders, government officials, lawyers, and academics to identify improvements in legal frameworks and policies that directly affect innovation. These include intellectual property, other R&D incentives, as well as industry-specific regulation in life sciences, information, and communications. CIP draws on the expertise of affiliated faculty across the University. A board of distinguished business leaders and former public officials advises the Center's leadership. For more information, visit law.duke.edu/innovationpolicy.

Introduction

From improving diagnosis and personalizing treatment decisions, to determining how best to meet the needs of underserved populations, artificial intelligence (AI) systems have the potential to revolutionize health care.[1] By 2021, the size of the health AI market will be about 11 times what it was in 2014, growing from $600 million to an estimated $6.6 billion.[2] This field is complex, and as with all technologies, not without risk. As such, it is important for manufacturers of AI-enabled software products to communicate information to clinicians, health system operators, and others about how to harness the benefits of AI while reducing risk.

AI refers to the ability of a machine to perform a task normally done by humans. AI-enabled clinical decision software is software that assists or automates the task of clinical decision-making around risk assessment, diagnosis, and treatment. AI-enabled software can be classified into two categories: rules-based and data-based algorithms. Rules-based algorithms use expert-derived rules and defined and logical processes to turn multiple inputs into an output—for example, an alert that reminds a physician that their patient is due for their colonoscopy based on clinically-accepted schedule guidelines. By contrast, data-based algorithms[*] are given sets of labeled input data (called "training data") and use programmed processes to derive relationships between the inputs and the so-called "labels"—for instance, labels that classify thousands of mammograms by whether or not the patient was eventually diagnosed with cancer. The derived relationships can then be used to predict how new input data is likely to be labeled. This paper will focus on data-based learning that uses labeled training data. This type of learning is generally called supervised learning (see **Figure 1**).

For years, providers have used rules-based AI in clinical decision software[†] to help make diagnoses and treatment decisions, manage population health, and carry out general administrative duties. However, recent advances in machine learning can improve the performance of software by opening the door to a range of new AI-enabled software that can guide more complex decision-making.

For logistical, technical, legal, or competitive reasons, manufacturers of AI-enabled tools, particularly data-based AI tools, might not disclose information about design, materials, and mechanism of action[‡] to regulators, purchasers, and users. Additionally, business considerations play a role in limiting disclosure. For-profit firms protect innovation through a variety of mechanisms, including patents and trade secrecy.[§] Incentives to keep training datasets proprietary may be reinforced by concerns about compliance with data privacy protections or security requirements. But risk assessment, and ultimately adoption, may be complicated if manufacturers or developers are reluctant to disclose trade secrets.

---

[*] Data-based AI is often referred to as "machine learning."

[†] This paper purposely uses a broad term *clinical decision software* to be inclusive of clinical decision support (CDS) software that is not under FDA authority, device CDS, and other Software as a Medical Device (SaMD) that goes beyond supporting a clinician in their decision-making by driving or automating the next medical intervention.

[‡] This white paper defines the term *mechanism of action* as a proven physiological explanation of how a medical product produces a therapeutic effect on a living organism or in a biochemical system.

[§] In addition to patents and trade secrecy, trademark law can play a role in protecting businesses against competitors that falsely claim that a given piece of software was developed by their firm. This paper does not address trademark law, as that law protects integrity of information about the *source* of a product, not information about how the product works.

This is because concerns around liability can be expected to influence health systems' evaluation of the associated benefits and risks of implementation and use.

Building and Testing Supervised Machine Learning Systems

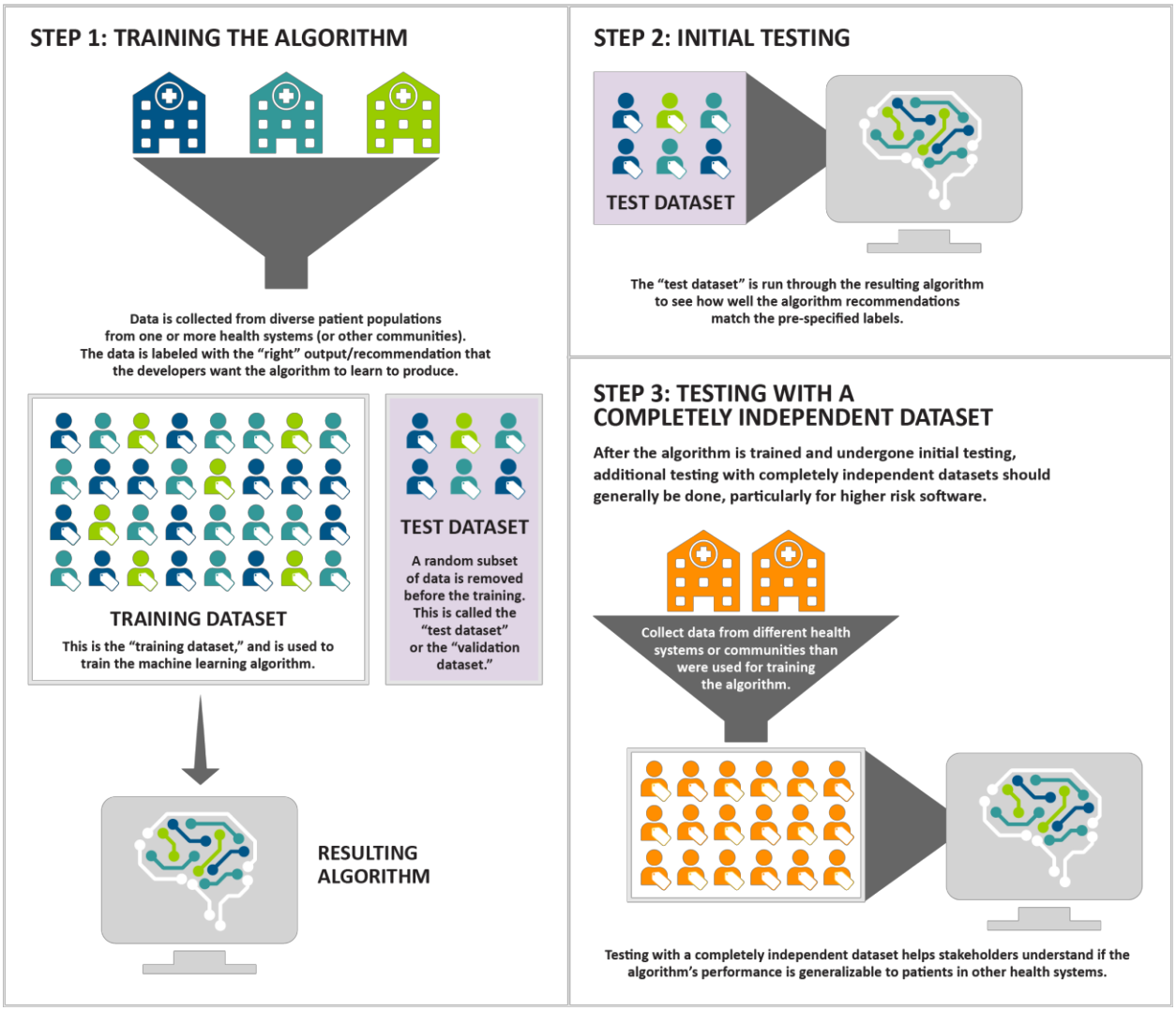

Figure 1. Building and testing supervised machine learning systems. These steps will be discussed in more detail starting on page 7.

This report explores how, in cases where certain information cannot be shared, alternative information could be used to satisfy stakeholder needs. The report is meant to serve as a resource for developers, regulators, clinicians, policy makers, and other stakeholders as they strive to develop, evaluate, adopt, and use AI-enabled medical products. We offer insight into how to incentivize innovation of safe and effective products while communicating information on how and when to use these products. Specific themes include the:

- Ways in which AI-enabled software in health care may differ from traditional medical products;
- Categories of information surrounding AI-enabled clinical software;

- Informational needs and governance structure around AI-enabled clinical software during the total product life cycle; and
- Role of regulatory incentives that protect developer investment, such as patents and trade secrecy, in information flow.

Discussion on informational needs and governance structure is also based on literature review, database searches, perspectives provided during meetings hosted by the Center of Innovation Policy at Duke Law and the Duke-Margolis Center for Health Policy, and individual stakeholder interviews.

What Makes AI-Enabled Clinical Decision Software Different from Other Medical Products?

AI-enabled software differs from traditional medical devices in important ways. These differences create challenges not only for regulators but also for clinicians, health systems, and others who may wish to adopt the technologies. For example, AI-enabled clinical decision software produces clinical recommendations but some of these AI-enabled products might not provide any information as to why and how those recommendations were reached. This lack of information may cause doubts in the minds of clinicians about whether the recommendations or decisions made by the software should be trusted.[3] Lack of trust can be exacerbated by the potential for clinician tort liability if the software recommendation is wrong.[4] Trade secrecy also may limit the amount of information that companies that develop software are willing to disclose, both about how these systems work and how they are built.

This section describes three key differences between AI-enabled software and other medical products: (1) software is powered by health data, which is heterogeneous, complex, and fast-changing; (2) software undergoes more rapid update cycles than other types of medical products; and (3) AI-enabled software might lack an explanation of "how it works."

Health Data

Traditional medical devices act on the structure or a sample of the body to produce results (although not through chemical action, which distinguishes devices from drugs). By contrast, software acts on health data which is inputted into the software and analyzed to come to a recommendation or prediction. In addition to acting on health data, machine-learning based software is also built with health data. Health data may consist of data produced through medical imaging, medical sensors such as electrocardiograms, or manually entered in electronic health records (EHRs) or other applications (see **Figure 2**). However, these data can be incomplete, inaccurate, or biased.[5] For example, information gathered from EHRs may be highly disparate in its accuracy and completeness, based on everything from different patients' socioeconomic status and potential language barriers to insurance documentation requirements and system workflows.[6]

Rules-based software produces more consistent output and generally uses limited, structured data elements as inputs. In contrast, data-based software often uses large, complex data sets as inputs; such data are more likely to reflect specific clinical workflows and the perspective of individual physicians, for example through the use of free text fields in EHR records. Patients' access to care, including tests, procedures and insurance coverage, also will affect the amount and types of health data available. Because health data is analyzed by software to reach recommendations, clear definitions around the data input requirements are necessary.

Data-based software manufacturers also use health data to build the algorithms used in their products. Training datasets should be examined to ensure that data is both reliable and relevant, in terms of both the population included and the metadata needed for accuracy and completeness. Due to the heterogeneity discussed above, software developed with data from one location (e.g., a region or hospital) may not work at other locations without significant changes to the software program. Bias can also be a concern. If a software system is not trained with sufficient data that originated from patients from ethnic minorities or patients with co-morbidities, or if the recorded data is historically biased because of socioeconomic status, race, or other criteria, the resulting software may not work well across all populations of patients and may even perpetuate existing biases within the health system.

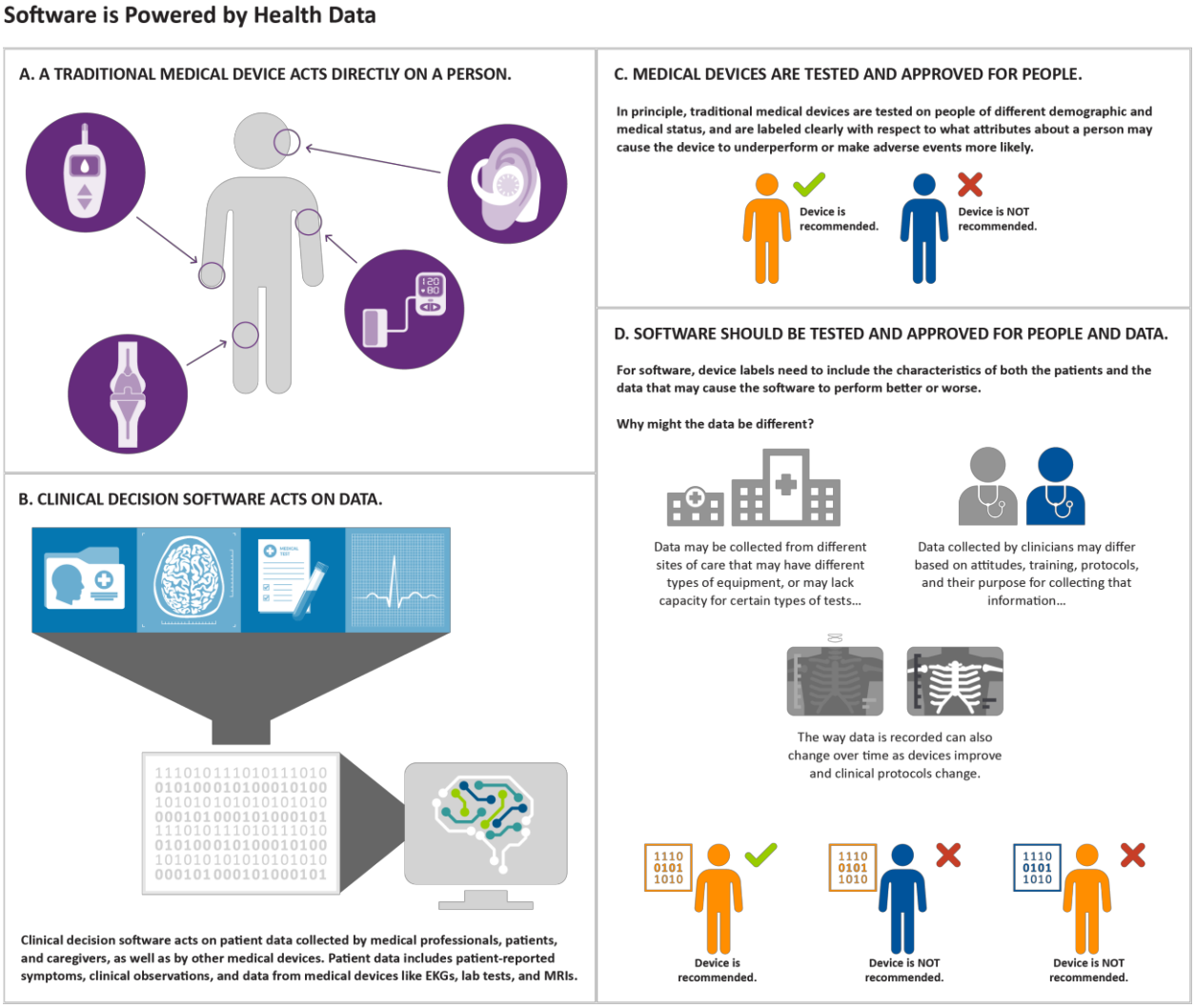

Figure 2. Software is powered by health data.

Accurate labeling of outputs—and the selection of accurate proxies (if proxies are used)—is also important. An October 2019 study described an algorithm that was trained using healthcare expenditures (cost) as a proxy to predict patients' level of risk of serious illness. Even after controlling for potential confounding factors, it was found that white patients use the health care system more than black patients, resulting in higher healthcare expenditures.[7] The algorithm assigned white patients

higher risk scores than black patients who were equally ill, thereby reducing the number of black patients who were identified as needing extra care by more than half.[8]

Health data also rapidly changes format and terminology over time as new clinical practices and medical products come into use. Even if these changes result in higher quality data that would improve human decision-making, software may need to be updated to interpret these changes or overall performance can suffer. So, while a well-maintained ultrasound imaging device will perform just as well (or poorly) several years after it is first used, software products designed to analyze ultrasound images may not perform as well when analyzing higher resolution images from a new type of imager or if protocols around the use of contrast agents changes. A software algorithm that uses data such as diagnoses, medication lists, etc., pulled from an EHR will likely degrade in performance over time if it is not updated to account for new medications, treatments, changing standards of care, and the way that these are documented. On the other hand, if (as discussed below) software is updated to reflect changes in underlying data, its performance can improve. As such, manufacturers, health systems and clinicians will need to work together to monitor system performance and update software as needed. The need for software to be regularly updated leads to the next key difference between clinical decision software and traditional medical devices.

Rapid Software Development Cycles

Rapid updating makes software distinctive among medical devices.[9] Manufacturers can act quickly to improve performance and correct problems found through real-world feedback by rapidly pushing updates to the users of those technologies. This is particularly true for AI-enabled software, as certain types of machine learning software have the potential to continuously update themselves in real-time (although it should be noted that clinical decision software of this type has not yet been approved or cleared by the U.S. Food and Drug Administration). These updates are critical to not just improving the product but also maintaining performance.

The rapid development cycle of software is a challenge for regulatory agencies, which have review and clearance processes based on more traditional devices with slower development cycles. The U.S. Food and Drug Administration (FDA) is working to adapt to these differences, including proposing a pre-certification program, which would be a voluntary pathway that would allow manufacturers and FDA to work together to enable rapid innovation and iterative improvements of clinical software while providing appropriate patient safeguards.[10,11] The FDA also released its "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning-Based Software as a Medical Device—Discussion Paper and Request for Feedback" in April 2019,[12] and asked for input from the public regarding how to meet the challenges in regulating the AI-enabled software.

Although frequent product updates should improve performance, they also present concerns for adopters and users of these devices. Best practices need to be developed to clearly inform software users of how updates may impact safe and effective product use. In addition, global updates (i.e., uniform updates sent to all installed software applications) may affect local performance in unexpected ways, emphasizing the need for regular performance monitoring.

Explainability

In the biopharmaceutical arena, certain popular therapies have unknown "mechanisms of action" or "modes of action."[13,14] This sort of uncertainty is less common with traditional medical devices, though

examples exist.[15] Some AI-enabled software products, however, may take uncertainty and its attendant risk—to yet a higher level.

Rules-based software is built on either clear physiological understanding or generally accepted clinical practice guidelines. Clinical practice guidelines may themselves have been built on observed statistical regularities rather than clear mechanism of actions. However, clinical guideline development is a multi-step and generally transparent process involving generating clinical evidence and drawing conclusions that are converted into clinical practice guidelines. Because users can be walked through the inputs and steps used to make the decision, backed up with clinically relevant guidance, rules-based software is generally considered to be "explainable." In contrast, certain data-based software products may not be able to provide stakeholders with a comprehensible explanation of how they weigh and combine inputs to come to a result, nor relate the recommendations back to physiological explanations. Because of this characteristic, this software is often referred to as "black box" software.[16]

All medical products, including AI-enabled software, can fail in unusual, unpredictable ways when the mechanism of action is not understood. In software that incorporates machine learning, failures may be partly due to unrecognized site-specific patterns or "clues" present in the training data, which can result in suboptimal performance when the system is deployed in new and different settings. For example, when researchers trained algorithms on pooled x-ray image data from sites with varying pneumonia prevalence, they found that the algorithms most likely used site-specific features in the images to significantly influence the resulting prediction, rather than simply relying on the underlying pathology. Because of these site-specific influences, the algorithmic models were not consistently generalizable to new health systems.[17]

When software is not explainable, rigorous performance testing can be performed to better understand the risks of the software. Prospective testing within the planned workflow is necessary to understand real-world product performance and whether the system may fail in unexpected ways. In addition, information provided at point-of-use, such as the certainty of a particular result, or what factors were weighed most heavily, may help users understand when to trust a particular result in the absence of a true explanation. The level of explanation or performance data necessary also can be calibrated as a function of risk posed by the software's intended use.

Categories of Information Surrounding AI-Enabled Clinical Decision Software

Categories of information exist that various stakeholders might want for AI-enabled clinical decision software products: how a software system fits into clinical workflow; what type of AI it is; how it was developed; how it works; and other information that may be useful to know about individual results (see **Figure 3**).

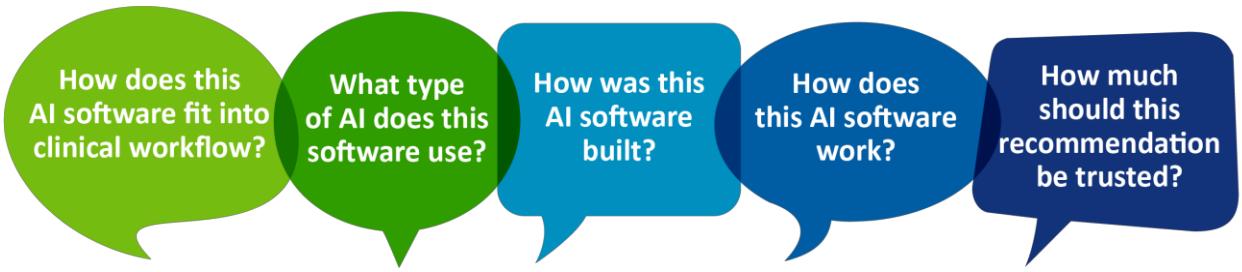

Figure 3. Categories of information for AI-Enabled Clinical Decision Software. *Various stakeholders throughout the total product lifecycle of a software product will want specific information of what the software does and how it fits into the workflow, what type of AI is used and how it was built, as well as information about how it works and when to trust the results.*

Information about the intended user of the software and how it relates to clinical decision-making should always be disclosed to all stakeholders. This baseline understanding should include the intended purpose and user of the device, and the significance of the result or recommendation to the user's clinical decision-making.

Relatedly, all stakeholders will need to understand whether an AI-enabled software product is designed to assist or automate a clinician's decision-making. If the software notifies a doctor of a possible medication interaction, or highlights certain areas of an x-ray for further review, the software is assistive and the final decision rests with the provider.

In contrast, autonomous AI-enabled software products diagnose or treat patients directly. This automatic action may occur through hardware that is part of the system, such as an implantable cardiac defibrillator that analyzes heart rhythm and sends an electrical shock to the heart. Alternatively, software may convey results to other users, who may not be trained to make the decision themselves, but who are still capable of taking next steps based on the results. It should be noted that the distinction between these two categories may not be clear cut, as there are multiple gradations in between.

To evaluate AI-enabled software, stakeholders also need to know what type of AI it uses. Is the software rules-based or data-based? If it is data-based, what learning algorithms were used to develop the software? Different types of algorithms are more suitable for different types of problems and data, similar to how certain statistical methods are more appropriate for certain types of analyses.[18]

Additionally, stakeholders need to know if AI-enabled software developed with machine learning is locked or continuously updating. Locked AI-enabled software means data-based techniques are used during development, but the software does not continuously learn and change over time. We are not aware of any continuously learning standalone software products that have been cleared or approved by FDA. However, continuously learning software products might be in use for administrative or population health purposes that are not under FDA authority.

Certain stakeholders might want more detailed information about the software development process as well. Full transparency for data-based AI could mean algorithmic transparency, which would include the code for the learning algorithm, as well as hyperparameters, training data, and other information needed to reproduce the algorithm(s) used in the software. For locked algorithms, transparency could also include model transparency—disclosure of the exact function or functions that are used to compute how all inputs are weighted and combined to produce the outputted recommendation. Stakeholders may also ask for detailed information about the training data, including how it was labeled.

As discussed further below, patents can, at least in theory, provide intellectual property protection even in the case of such full transparency.[**] However, difficulties in enforcing patents, and a desire on the part of some patent applicants to attempt to maintain both patent and trade secrecy protection over the same information, may make applicants reluctant to provide full transparency. Recent U.S. Supreme Court patent eligibility cases have made patenting of both medical diagnostics and software more difficult. When companies do not have secure patent protection, they may rely even more heavily on trade secrecy to protect their investment in innovation.

One context in which trade secrecy may be particularly important is training data. Patents and copyright do not extend to raw data. The restrictions on information flow required by trade secrecy law may also align with privacy-related legal prohibitions against disclosing training data that contains personal health information (PHI).

However, even if manufacturers are reluctant to disclose training datasets, summary information on patient populations represented, including demographics, social determinants of health, geographical region, comorbidities, and genetic markers, will still be useful. Any data curation, inclusion/exclusion criteria for adding patient data to the training dataset, and clear methodologies for how the data was labeled should also be part of this summary, and be incorporated into device labelling. Summary information on patient populations used for training should shed at least some light on potential biases and on whether the training population resembles the patient population of interest to the stakeholder.

A common question that stakeholders have regarding novel medical products is: how does the product work? For traditional medical devices, information regarding how to reproduce the device (of the sort that should be disclosed in patents) should also provide insight into how the device works. Unfortunately, in the case of data-based software, information required to reproduce the algorithm driving a software product may not be helpful for human understanding of what that software is doing.

A true explanation delineates exactly how the software product will process input data to produce a result. Software that utilizes rules-based AI can always give "true" explanations, and certain types of machine learning or product designs also can provide some explainability. For black box algorithms,

[**] Patent doctrine requires that the information disclosed in the patent provide the basis for reproduction—specifically, that it shows "one skilled in the art" how to make and use the claimed invention.

statistical techniques that can produce a "likely" explanation are also being explored.[19]

Because of this limited explainability, detailed performance data should become more important to stakeholders. Indeed, rigorous evidence around performance should be required by all stakeholders, regardless of the type of AI used (although requirements on rigor may differ based on the risk posed by the AI). It is therefore critical for stakeholders to clearly communicate what type of performance data is being asked for and given. For example, studies involving data-based AI should include information regarding whether performance results are coming only from a validation dataset that was separated from the original training data before training began, or from a completely independent dataset collected from a different source and/or at a different time.[20] Testing on a completely independent dataset will shed light on whether software performance depends on data features or patterns specific to the sites from which the training data was collected.

Stakeholders also should understand whether testing was retrospective or prospective, and whether the product was tested in the environment and within the workflow in which it is intended to be used.[21] A 2017 JASON report on AI for Health and Health Care recommends that rigorous procedures be developed for approving and accepting AI-enabled software into clinical practice, including testing and validation approaches for AI algorithms to evaluate performance under different conditions.[22]

Furthermore, adopters need information on how software inputs should be structured and defined. For example, does the software only work with images from particular manufacturers or models of imaging equipment? Having this information will enable stakeholders to understand if their own data can be used effectively by the software. Prior to adoption, potential adopters may also want to consider testing the software on their own data to evaluate local performance.

Finally, clinicians need appropriate information at the point-of-use about software system results to determine how heavily to weigh them in their decision-making. This information can include the certainty of the software for a specific result or the key input features that led to a specific recommendation. Users also may find it useful to have information about whether their patient significantly differs demographically or medically from the training and testing population. It is important that software systems be designed to communicate such information quickly, and in readily understandable ways to accommodate clinicians' busy schedules. However, FDA has cautioned against using labeling beyond what is typical in clinical settings. Information provided should be in line with the labeled use of the product and, for automated systems, information that users are not trained to interpret should be avoided as is may be counter-productive.

Governance Structures for Information Flow Across the Total Product Lifecycle

Once we understand the categories of information surrounding AI-enabled software, it is important to understand the regulatory and institutional frameworks that govern how this information might be requested or supplied by stakeholders at each point of the total product lifecycle (development, regulation, adoption, monitoring, and use). The following sections address governance issues surrounding information flow. The discussion is based on literature review, database searches,

perspectives provided during meetings hosted by the Center of Innovation Policy at Duke Law and the Duke-Margolis Center for Health Policy, and individual stakeholder interviews.

Development

Patent Law

Patent law requires applicants to provide a disclosure that enables scientists of "ordinary skill" to "make and use" the invention. Relatedly, applicants must provide a "written description" about the structure of the invention they are claiming. Under the patent system, this disclosure, which mirrors the scientific research and publication norm of reproducibility, is the *quid pro quo* that inventors provide to society in exchange for a time-limited right to control both direct competition and cumulative innovation in their area of invention.

While disclosure through patents can occur, the U.S. Patent and Trademark Office does not always enforce disclosure requirements. Moreover, applicants often file for patents early in the R&D process, before a full understanding of the invention has been achieved.[23]

Legal and ethical challenges unique to AI-enabled health software may impede disclosure by developers. First, software patent law is highly unsettled (as mentioned earlier), so companies might not feel confident in the protections that patents otherwise confer. Second, training data might contain personally identifiable information or information that could be combined with other data to re-identify the individuals who were the source of that data. The potential for identification (or re-identification) raises privacy concerns, including potential violations of the Health Insurance Portability and Accountability Act (HIPAA) depending on the type of data used. Finally, in the case of AI-enabled software, although reproducibility allows some scientists to have confidence in the veracity of their results, this does not mean that the model is comprehensible to all scientists, let alone to other stakeholders.

To investigate patent disclosure further, we examined the patents associated with several prominent data-based AI software products recently cleared by the FDA.[24] These included the QuantX software for reading MRIs to detect abnormalities suspicious for breast cancer; the Viz.AI ContaCT device for detecting, and triaging, suspected large vessel occlusions in an emergency room context; and the IDx-DR software for analyzing retinal images to provide a primary care physician with a recommendation regarding whether diabetic retinopathy had been detected. We found that the patent disclosures associated with these products contained at most only a brief, highly general, discussion of training data, the training process, or criteria used for validation.

Funding

We also examined the issue of venture capital (VC) funding, particularly in light of Supreme Court decisions that make patenting of AI-enabled clinical decision software more challenging. Our data indicate that the Court decisions have not deterred VC investment. To the contrary, as with AI-enabled health generally,[25] VC investment in AI-enabled clinical decision software has risen in recent years.[26] That said, one of the venture capitalists we interviewed did indicate that greater ability to patent would further increase investment in small machine learning firms. Each of the venture capitalists we interviewed viewed developer secrecy over training data and model details as key mechanisms for protecting investment in innovation.[27]

FDA Regulation

FDA defines medical devices as instruments used in the diagnosis, cure, mitigation, treatment, or prevention of disease, that can affect the structure or function of the body through non-chemical means. This definition includes certain types of software, termed "Software as a Medical Device (SaMD)" that are "intended to be used for one or more medical purposes and to perform these purposes without being part of the hardware of the medical device."[28] Thus far, AI-enabled SaMD have been cleared under either the 510(k) pathway for devices substantially similar to other devices on the market or through a de novo classification for novel low-to-moderate risk devices. FDA has published multiple documents on SaMD. These papers include discussion of both CDS and AI-enabled SaMD, and have not suggested that there will be systematic differences in how AI-enabled software will be evaluated relative to other software. Below, we review these documents to understand the types of information that that FDA may request from manufacturers as part of regulatory review.

As an initial matter, it is important to note that not all clinical decision software is SaMD. Under the 21st Century Cures Act (Cures Act) of 2016, software that is not SaMD includes software that presents institution-specific best practices, facilitates access to treatment guidelines, or software that acts in an administrative or quality improvement capacity. The Cures Act also establishes a somewhat complex scheme for determining what types of clinical decision support (CDS) software are, and are not, subject to FDA authority.[††]

Since 2016, FDA has worked to interpret the CDS software provisions of the Cures Act. In September 2019, FDA released an updated draft of its CDS Software guidance, which removes certain types of CDS software from FDA authority (FDA calls these "non-device CDS software").

The Cures Act specifies several criteria that determine whether a CDS software product is a medical device and therefore under FDA authority. The first criterion relates to the required input data. Any product that uses a "medical image or signal from an in vitro diagnostic device, or pattern or signal from a signal acquisition system" as an input remains under FDA authority ("device CDS software").

Even if the CDS software does not use imaging or signal data, it must pass additional tests in order to fall outside of FDA authority. The software should be intended for the purpose of "displaying, analyzing, or printing medical information about a patient" in order to support or provide recommendations "to a health care professional about the prevention, diagnosis or treatment of a disease or condition". Further, it must allow the "health care professional to independently review the basis for recommendations that software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient."

The "independent review" criterion has proved especially challenging to interpret. In an example from the appendix to its 2019 guidance, FDA states that software developed with machine learning could meet this criterion if "the logic and data inputs for the algorithm and the criteria for [the

[††] In its 2019 draft CDS guidance, FDA limits the term clinical decision support (CDS) to software "supporting or providing recommendations to an [healthcare professional], patient, or caregiver." Software that drives or automates diagnosis or treatment decisions would be considered a medical device, but not CDS software. The term "clinical decision software" used in this paper is intentionally broad, to encompass both device and non-device CDS, as well as software that drives or automates diagnosis or treatment.

recommendations] were explained and available to the [health care professional]." The FDA's statement suggests that AI-based software that is not able to provide a human-comprehensible explanation for how the software works or details about specific recommendations will remain under FDA authority. As for FDA's reference to data inputs being available to the user, it is unclear if machine-learning software that uses large numbers of input elements will be able to comply with that requirement in a way that reasonably allows independent review.

For CDS software products under FDA authority, FDA will use a risk-based approach and take into consideration four factors (see **Table 1**):

1. The significance of the software result in clinical decision-making;
2. The clinical context of the health care situation;
3. The type of user (health care professional versus patient or caregiver); and
4. The ability of that user to independently review the basis for the recommendation.

The guidance states that the International Medical Device Regulators Forum (IMDRF) framework will be used to assess the first two factors.

Table 1. Summary of regulatory policy for CDS software functions. Modified from FDA's 2019 CDS draft guidance document.[‡‡]

CDS software that "informs" clinical decision-making	User: Healthcare Providers		User: Patient/Caregiver	
	Can Independently Review	Can Not Independently Review	Can Independently Review	Can Not Independently Review
Non-Serious	Not a medical device	Enforcement discretion	Enforcement discretion	Oversight focus
Serious	Not a medical device	Oversight focus	Oversight focus	Oversight focus
Critical	Not a medical device	Oversight focus	Oversight focus	Oversight focus

Because FDA will incorporate these four factors (the IMDRF categories regarding clinical context and the significance of the information, the ability to independent review recommendations, and the intended user) into their risk assessment, clear labels would be extremely useful. The intended user is already a prominent part of the label, but the specific clinical context and the significance of the information as defined in the guidance are often less clear.

Beyond interpreting the Cures Act, FDA is also examining the possibility of a new regulatory model for software. Announced in July 2017,[29] the Software Pre-Certification ("Pre-Cert") Pilot Program is meant to help inform "the development of a future regulatory model that will provide more streamlined and efficient regulatory oversight of software-based medical device."[30] The program is being developed in response to FDA's recognition that its traditional approach to regulating medical devices "is not well

[‡‡] According to a November 2019 FDA webinar, "enforcement discretion" indicates that, at this time and based on FDA's current understanding of the risks of these devices, FDA does not intend to enforce compliance with applicable device requirements.

suited to the faster iterative design, development, and type of validation" used for many SaMD products.[31]

The latest working model, updated in January 2019, explains that companies that have met the pre-certification qualifications will still go through a review pathway determination for individual software products, based on the risk of the software. The working model lists product-level elements that may contribute to the review pathway determination, which include "an explanation of how the software works" as well as "instructions and limitations on use" and the "critical features/functions of the SaMD that are essential to the intended significance of the information" to decision-making.[32] FDA lists the clinical algorithm as one of the product-specific elements in the streamlined review process description, including mechanism of action (although the term is not defined). As such, it is currently unclear whether a lack of explainability will affect the risk assessment of a software product under Pre-Cert or how the review pathway selected might change if some of those product-level elements are absent. As the pilot continues, more details might be shared to clarify these questions.

The Pre-Cert model states that "validation of the clinical algorithm is of primary importance and would be fully described and would include both protocols for testing and results demonstrating performance." In the April 2019 discussion paper on AI/ML and continuous learning, FDA emphasizes the need to have "large, high-quality, and well-labeled data sets" to have a robust algorithm.

Stakeholder interviews suggest that FDA is not asking companies for full training data or detailed information about the algorithms for AI-enabled software products developed with machine learning. Instead, companies have shared summary information on training data and have provided more detailed information on clinical study data, methodology, and results for FDA assessment. Our interviews indicate that although companies are not averse to disclosing details regarding the underlying model to regulatory agencies, they are hesitant to hand over detailed training data that they view as a trade secret.

Manufacturers also report that FDA is interested in understanding the user experience. Relevant information includes the amount or type of information the end user receives as well as the significance of the software recommendation and fit for the end user.

Manufacturers in our stakeholder interviews report their experiences with FDA as positive. Stakeholders view FDA's proposed Pre-cert Program as a positive sign that FDA is thinking deeply about how regulatory policy should change to foster new innovations while maintaining patient safety. To keep up with device review, manufacturers also acknowledge the need to increase FDA's ability to recruit and retain relevant talent or expertise. FDA has already released statements about its efforts to increase the "number and expertise of digital health staff at FDA"[33] and to create partnerships with medical product centers, academic stakeholders, and other partners in order to "improve the ability of FDA reviewers and managers to evaluate products that incorporate advanced algorithms and facilitate the FDA's capacity to develop novel regulatory science tools."[34]

Stakeholders also agreed that adopters view FDA clearance/approval as a positive indicator of efficacy. We will discuss adoption more in the next section.

Adoption and Use

The next step in the product lifecycle is two-fold: the adoption and implementation of a software product into a provider system, and the decision by a health care provider and patient at point of use to incorporate the software recommendation into their decision-making. Convincing healthcare systems to adopt and use AI-enabled software will depend on the software's perceived or demonstrated ability to improve health outcomes, the costs and financial benefits seen by adopters of the technology, how well it can be integrated into clinical workflows, alignment to the standard of care, and the relevant law and regulations on tort liability.

Depending on where the product is deployed, different types of information may be required for health systems, providers, and patients to trust these technologies enough to adopt and use them. [35] For initial adoption decisions, data showing the software can improve the system's overall patient population health may be an important factor. At point of use, however, the certainty of the output or the logic and key inputs leading to the recommendation may be more important to clinicians, as well as what medical or other patient factors may affect the accuracy of the recommendation. Such information can help them discern when specific recommendations may be more or less relevant for the particular patient in front of them.

Adoption

Multiple stakeholders mentioned that FDA approval or clearance was a helpful mark of quality and effectiveness. However, stakeholders also used other information when making a decision to adopt a software product. Our interviews suggest that decision-makers are most interested in information pertaining to performance, including sensitivity and specificity analyses.[§§] Health systems may also ask for explanations on how the software works and will improve day-to-day clinical processes.

Provider system stakeholders also spoke about the need for guidelines and systems to properly assess new AI-enabled products. A user guide released in November 2019 delineates how provider systems should evaluate diagnostic products developed with machine learning.[36] The authors recommend starting the assessment with a determination that the machine learning method is appropriate giving the function of the resulting software and the type and amount of data used to train the algorithm. The number and regularization of parameters should also be assessed to determine if overfitting[***] may be a concern.

Next, the algorithm should be validated and the validation methods should be examined. Was the validation dataset completely separate from the datasets used to train and tune the algorithm? Is the reference standard high quality? This latter question can be a challenge when there is no gold standard for comparison.[37] With results that seem "too good to be true" or if unexpected associations or correlations are found, the performance can be validated in additional patient cohorts to "ensure that the results are not due to artifacts in the machine learning systems, confounding factors, or flaws in the study design."[38] Repeatability and reproducibility of the software recommendations should also be

[§§] Sensitivity and specificity analysis are generally used in medical diagnosis to determine the ability of a test to correctly identify the true positive rate or those with a disease (also known as sensitivity) in addition to the true negative rate, or the ability of the test to correctly identify those without the disease (also known as specificity).

[***] Overfitting occurs when an algorithm is built to match the training data too closely. Because the algorithm creates rules for "noise" that is only present in that specific dataset, the software is not generalizable to other data sets.[39]

assessed, by examining how small changes in input data affects the outcomes, as well as how input data from different hardware, operators, and protocols may affect real-world performance.

One of the concerns frequently discussed in academic circles is whether health care should demand "explainable" software from developers or if "black box" software is acceptable. AI enthusiasts commonly say many clinicians (including themselves) use medical devices daily that they don't understand. However, such use generally occurs with the knowledge that many experts (such as the manufacturers, FDA, and possibly even a technical assessment committee within the provider system) do understand how the device works and can test for potential complications based on that understanding. This is different from black box algorithms, where there is no explanation that even experts can understand about how the software is analyzing the input data to come to outputs.

It should be acknowledged that it is not unusual for the mechanism of action of a medical product to not be fully understood, and this is generally addressed with rigorous testing to alleviate concerns with safety and efficacy. In an editorial published in January 2020, researchers argue that a practical solution could be to demand different levels of explainability based on use case and the balance of benefit and risk.[40] The authors also recommend rigorous performance studies and local pilot testing before and after implementation if adopting "black box" software.[41]

Almost all the stakeholders interviewed stressed that AI-enabled clinical decision software can enhance workflows, positively influence care decisions, and improve outcomes.[†††] In order to achieve these goals, information must flow in both directions. Provider systems can help developers by being more open about their processes and needs, while developers can bring in people who are well-versed in these systems to help consult during the development process. In addition, best practices are needed on how developers can efficiently provide evidence of improved workflow and outcomes, or take on risk in value-based outcome arrangements with provider systems when there are questions regarding how much realized value will be gained by the patients or system.

Even with evidence of clinical utility, stakeholders recognized, as previously discussed, multiple factors that might affect whether a specific software will work effectively with a particular health system's patient population, data systems, and workflow. Therefore, some of the interviewed provider systems test all algorithms with data from provider system patient populations before making a final decision to adopt a system. However, even those who did testing noted that not all health care systems have adequate resources for testing.

Stakeholders repeatedly mentioned algorithmic performance may degrade over time due to the ever-changing nature of input data. This makes the ability to continually monitor the performance of the algorithm critical. Despite concerns, none mentioned systematic processes outfitted to do this type of monitoring. Additionally, despite increasing interest in AI-enabled software, machine-learning systems have not yet achieved widespread use in health care systems. A January 2020 Technology Review Insight survey found only 10 percent of health care institutions have deployed one or more AI applications, with another 17 percent having deployed one or more AI pilot projects.[42] However, another 45 percent of the institutions surveyed are in the process or are planning to deploy AI in the next 2 years. Notably, of the institutions that use or plan to use AI, 74 percent plan to develop their own customized AI algorithms.

[††††††] In fact, the standard of care could evolve to require that the performance of the provider by augmented by software.

While ensuring that algorithms are trained on appropriate patient population and workflows is key, customization leaves the institution with the sole responsibility of monitoring performance over time and updating the software as needed. For healthcare systems that are very large, there may need to be customization within the system itself based on the location in which the product is deployed.

Minimizing security and privacy risks through proper controls or data governance also will be key to dynamic health system operationalization. Interviewees stressed implementing processes in a way that seamlessly integrates with the user experience and fosters patient trust by safeguarding vital information.

Clinician Acceptance and Use

More than one stakeholder interview transitioned into a conversation about workforce training and the user experience. As hospitals strive to cultivate AI systems and continue to evolve, algorithms should relay critical information about individual recommendations to clinicians in a user-friendly manner. However, clinicians also need to be provided ample opportunity to understand basic foundational concepts. The April 2019 European Commission Ethics Guidelines for Trustworthy AI speak to this need for human agency and oversight stating that users should be "given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system."[43] The June 2018 American Medical Association (AMA) policy on the use of augmented intelligence (also known as AI) in health care underscores the need for thoughtfully designed, high-quality, and clinically validated AI-enabled software, and the ability for the provider to "understand AI methods and systems sufficiently to be able to trust an algorithm's predictions."[44] However, without a proper knowledge base, clinicians might not be able to effectively work with, or manage, the AI system, or know which specific questions they can ask to gain appropriate insight.

As a first step, clear and accessible product labeling for clinicians to refer to during use can be crucial to allowing the user to fully assess risks and biases that might arise from the algorithmic training process.[45] Medical device labels include information on the benefit-risk profile as well as indications for proper use. Stakeholders familiar with the FDA approval process proposed that key elements might include aggregated stats about the patient population used to train the model (demographics of training and validation) in addition to information about accuracy of the algorithm tested on completely independent validation sets.

While label information is important, stakeholder discussions also revealed that information about specific recommendations might be more valuable to busy clinicians at the point of care. Suggestions included incorporating information about how a software's key input factors influencing the recommendation and information to help clinicians understand how many "patients like theirs" are included in the training data. Stakeholders agreed that default information should be limited and quick to digest visually rather than requiring users to scroll through dense text, however some suggested that users should be able to "click" to get more detail.

The type, level of software autonomy, and degree of information provided to clinicians will affect the amount of liability they might be willing to accept (especially when the results given by the AI system differ from their own clinical judgement). Some stakeholders indicated that users (and health systems)

are hesitant to take on undue risk and invest in AI-enabled systems without fully understanding how liability will play out in the long term and with whom the responsibility lies.

The authors of an October 2019 study examine possible scenarios in which the recommendation from the software does or does not differ from standard of care, the clinician follows or rejects the recommendation, whether the patient outcome is good or bad, and what the potential liability may be in each case. Their work suggests that, at least until the use of medical AI itself becomes part of the standard of care, "the 'safest' way to use medical AI from a liability perspective is as a confirmatory tool to support existing decision-making processes, rather than as a source of ways to improve care."[46] The authors recommend that clinicians protect themselves by gathering information and asking clinical societies to develop best practices in how to evaluate both a new AI product overall and individual recommendations from that product, as well as ensuring that products have been thoroughly vetted before procurement. Furthermore, the authors suggest that physicians should ask questions from their malpractice insurers about use of AI-enabled clinical decision software. Changes might be required to both coverage contracts and liability laws as AI-enabled software becomes more widespread.

Patient Acceptance

The data on patient acceptance are mixed. A September 2019 survey revealed that about 45 percent of respondents said they were interested in their physician using AI to help with a diagnosis, due to the potential for a more accurate diagnosis, a reduction in human error, and/or faster treatment decisions.[47] However, a May 2019 paper found that patients were less likely to use or pay for a service if the health care was provided by an AI system instead of a human provider.[48] Although these patients did not believe the AI provided inferior care, the patients were skeptical that the AI was able to provide care that was tailored to their circumstances and unique patient profile.[49]

The amount or level of information patients want can differ based on whether the software is assistive or automated, and might affect how willing they are to embrace certain technologies. Typically, patients want AI that assists clinicians as opposed to automating them—acting as a complement instead of a replacement, especially with sensitive treatments or lasting interventions.[50] A January 2019 study conducted by Deep Mind and RSA revealed that increased ease of understanding with respect to information conveyed to the patient does not necessarily translate into increased levels of trust. Of those respondents surveyed, 36 percent were likely to support automated AI systems if they were able to request an explanation of the steps or processes it took to come to the decision, with only 20 percent indicating increased support of the technology if it were explainable to an individual with no technical expertise.[51]

Conclusion

Stakeholders require substantial information about AI-enabled software to effectively harness its benefits and mitigate risk. Some information regarding AI-enabled software is comparable to information stakeholders need to know about traditional medical products. However, AI-enabled software can present additional informational demands. Moreover, unique business concerns and technical challenges may at times create mismatches between information regulators and adoptersdesire and information developers are willing or able to provide. Our work examined where these mismatches may exist and what information regulators and adopters of AI-enabled software may accept in lieu of traditional information.

Our research suggests that, as an empirical matter, conflicts over trade secrecy have not been a significant issue so far. We found that regulators and adopters' informational needs vary based on how recommendations produced by AI-enabled software are used in clinical decision-making, as well as the clinical context. Furthermore, for the moment, regulators and adopters' expectations align with the amount of information currently disclosed by manufacturers. Manufacturers are reluctant to share training data or disclose details of trained models, but they are generally willing to share summary information on both. Thus far, stakeholders have not been pushing for more detailed information.

In part, the current congruence of stakeholder expectations may arise because most products used today are low- to medium-risk. As more autonomous and higher risk products that may require more trust emerge, expectations could diverge and tensions arise. For example, given developers' reluctance to share training data and full model details with third parties, including the FDA, high-risk scenarios where access to such information was important may create tensions. More troubling is the possibility that disputes have not yet arisen because, even now, adopters are asking for insufficient information. For example, in contravention of emerging best practices, some adopters do not appear to be asking for performance data gleaned from a dataset collected completely independently from the initial training dataset.

Our research also shows that basic education about AI-enabled products is necessary for stakeholders, particularly end users, to understand the type of information they need to safely use AI-enabled clinical decision products. Policy makers, hospital systems, and researchers will need to work together to provide end users with educational resources that promote understanding the information needed during their decision-making process. Currently, FDA is working on expanding and fortifying its workforce through active recruitment efforts. Hospital systems need to consult with clinicians and internal technological assessment committees to create systematic plans for evaluating products and educating their workforce. And though there has been an increase in literature on how to effectively evaluate AI-enable products in health care, it might be useful for a centralized third-party to act as a repository for these evaluations—although this will not account for challenges around site-specific data and workflow issues.

Below are initial recommendations on information that should be shared as stakeholders explore, evaluate, adopt, use, and monitor emerging AI-enabled products:

- Provider systems should be open about their internal process challenges and informational needs so manufacturers are better able to develop products that solve real problems and fit into the health system work flow. In parallel, manufacturers need to bring in experts who are well-versed in health system workflows and curating products for the user experience. Manufacturers also need to show evidence of the clinical utility of their product, not just the accuracy of the results.
- As products emerge that have a higher risk profile, procedures should be developed by which information considered by developers to be a trade secret (e.g. training data and model details) may need to be shared with trusted third parties (e.g., the FDA) that can evaluate the information.
- Conveying performance data on an independent test set, information regarding the certainty of the recommendations, and, if technically possible, key weighted factors in the algorithm's

decision-making process can increase stakeholder trust as they evaluate the product and determine whether to adopt or use it.

- Information about the intended use (such as the purpose, user, significance of decision, level of autonomy given, patient population) should always be disclosed publicly, in addition to summary information about the training data, labeling methodology, and testing or validation process.
- Manufacturers need to clearly define data input requirements, including the structure and definition of each data element, so adopters can understand if the algorithm can be used effectively with their patient population and workflows. Defining the expected clinical context of the data collection may also be important.
- Stakeholders should develop a set of best practices and recommendations on how to best evaluate a new AI-enabled software product, including guidelines for how to thoroughly vet products before procurement.

Finally, because AI-enabled software can fail or break in unexpected ways, manufacturers and health systems should work together to monitor system performance after implementation, including updating as needed, and share information about product limitations and adverse or near-miss events.

AI has the potential to streamline workflows, increase job satisfaction, reduce spending, and improve health outcomes. A 2020 survey demonstrates that 89 percent of healthcare executives believe that AI is already creating efficiencies in health systems, and 91 percent believe it has the potential to increase patient access to care.[52] Estimates also show that AI can help address about 20 percent of unmet clinical demand.[53,54] However, to achieve these goals responsibly and cultivate long term success, ensuring that the right information is shared with the right stakeholder at the right time will be essential.

References

[1] Bresnick, J. (2018). "Top 12 Ways Artificial Intelligence Will Impact Healthcare." *Health IT Analytics.* Retrieved from https://healthitanalytics.com/news/top-12-ways-artificial-intelligence-will-impact-healthcare

[2] Accenture. (2017). "Artificial intelligence: healthcare's new nervous system." Retrieved from https://www.accenture.com/t20171215T032059Z__w__/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf

[3] Duke-Margolis Center for Health Policy. (2019). "Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care." Retrieved from https://healthpolicy.duke.edu/sites/default/files/atoms/files/dukemargolisaienableddxss.pdf

[4] Sullivan, H.R., & Schweikard, S.J. (2019). "Are current tort liability doctrines adequate for addressing injury caused by AI?" *AMA J Ethics.* Retrieved from https://journalofethics.ama-assn.org/article/are-current-tort-liability-doctrines-adequate-addressing-injury-caused-ai/2019-02

[5] Duke-Margolis Center for Health Policy. (2018). "Characterizing RWD Quality and Relevancy for Regulatory Purposes." Retrieved from https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf

[6] Gianfrancesco, M.A., et al. (2018). "Potential biases in machine learning algorithms using electronic health record data." *JAMA Intern Med. Retrieved from* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/

[7] Obermeyer, Z., et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." *Science.* Retrieved from https://science.sciencemag.org/content/366/6464/447

[8] Ibid.

[9] FDA. (2019). "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/122535/download

[10] FDA. (2017). "Digital Health Innovation Action Plan." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/106331/download

[11] FDA. (2018). "Developing Software Precertification Program: A Working Model." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/113802/download

[12] FDA. (2019). "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/122535/download

[13] Jilani, T.N., & Sharma, S. (2019). "Trihexyphenidyl." *StatPearls.* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK519488/

[14] Rosenbaum, S.B., & Palacios, J.L. (2019). "Ketamine." *StatPearls.* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK470357/

[15] Conway, C.R., & Xiong, W. (2018). "The Mechanism of Action of Vagus Nerve Stimulation in Treatment-Resistant Depression: Current Conceptualizations." *The Psychiatric Clinics of North America.* Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30098653

[16] Duke-Margolis Center for Health Policy. (2019). "Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care." Retrieved from https://healthpolicy.duke.edu/sites/default/files/atoms/files/dukemargolisaienableddxss.pdf

[17] Zech, J.R., et al. (2018). "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study." *PLOS Medicine.* Retrieved from https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

[18] Crown, W.H. (2015). "Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions." *Value in Health*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1098301514047913

[19] Turek, M. (2018). "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency.* Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

[20] Liu, Y., et al. (2019). "How to Read Articles that use Machine Learning User's Guide to the Medical Literature." *JAMA.* Retrieved from https://jamanetwork.com/journals/jama/fullarticle/2754798

[21] Topol, E. (2019). "High-performance medicine: the convergence of human and artificial intelligence." *Nature Medicine.* Retrieved from https://www.nature.com/articles/s41591-018-0300-7

[22] JASON. (2017). "Artificial Intelligence for Health and Health Care." *The MITRE Corporation.* Retrieved from https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf

[23] Cotropia, C.A. (2009). "The Folly of Early Filing in Patent Law." 61 Hastings L.J. 65.

[24] Rai, A.K., et al. (2020). "Accountability, Secrecy, and Innovation in AI-Enabled Clinical Decision Software." *Journal of Law and the Biosciences.* In press.

[25] Ibid.

[26] Ibid.
[27] Ibid.
[28] IMDRF SaMD Working Group. (2013). "Software as a Medical Device (SaMD): Key Definitions." *IMDRF.* Retrieved from http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf
[29] FDA. (2017). "Digital Health Innovation Health Plan." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/106331/download
[30] FDA. (2019). "Developing Software Precertification Program: A Working Model (v1.0)." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/119722/download
[31] FDA. (2017). "Digital Health Innovation Health Plan." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/106331/download
[32] FDA. (2019). "Developing Software Precertification Program: A Working Model (v1.0)." *U.S. Department of Health and Human Services.* Retrieved from https://www.fda.gov/media/119722/download
[33] Gottlieb, S. (2018). "Transforming FDA's approach to digital health." Retrieved from https://www.fda.gov/news-events/speeches-fda-officials/transforming-fdas-approach-digital-health-04262018
[34] Gottlieb, S. (2019). "The Role of Real-World Evidence in Regulatory and Value-Based Payment Decision-Making." *Remarks made at Bipartisan Policy Center.* Retrieved from https://bipartisanpolicy.org/events/the-role-of-real-world-evidence-in-regulatory-and-value-based-payment-decision-making/
[35] Duke-Margolis Center for Health Policy. (2019). "Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care." Retrieved from https://healthpolicy.duke.edu/news/white-paper-release-current-state-and-near-term-priorities-ai-enabled-diagnostic-support
[36] Liu, Y., et al. (2019). "How to read articles that use machine learning." *JAMA*. Retrieved from https://jamanetwork.com/journals/jama/article-abstract/2754798
[37] Adamson, A.S., & Gilbert Welch, H. (2019). "Machine learning and the cancer-diagnosis problem—no gold standard." *NEJM*. Retrieved from https://www.nejm.org/doi/full/10.1056/NEJMp1907407
[38] Liu, Y., et al. (2019). "How to read articles that use machine learning." *JAMA*. Retrieved from https://jamanetwork.com/journals/jama/article-abstract/2754798
[39] Hanelman, G.S., et al. (2029). "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods." *AJR*. Retrieved from https://www.ajronline.org/doi/full/10.2214/AJR.18.20224
[40] Wang, F., et al. (2020). "Should health care demand interpretable artificial intelligence or accept "black box" medicine?" *Annals of Internal Medicine.* Retrieved from https://www.acpjournals.org/doi/10.7326/M19-2548?searchresult=1
[41] Ibid.
[42] MIT Technology Review Insights. (2019). "The AI effect: how artificial intelligence is making health care more human." Retrieved from https://mittrinsights.s3.amazonaws.com/ai-effect.pdf
[43] High-Level Expert Group on Artificial Intelligence. (2019). "Ethics guidelines for trustworthy AI." Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Human%20agency
[44] American Medical Association. (2018). "Augmented intelligence in health care: report 41 of the AMA Board of Trustees." Retrieved from https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/5b6aed0a758d4610026a719c/1533734156501/AI_2018_Report_AMA.pdf.
[45] Geis, J.R., et al. (2019). "Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement." Retrieved from https://www.jacr.org/article/S1546-1440(19)30944-5/pdf
[46] Ibid.
[47] UnitedHealth Group. (2019). Retrieved from https://www.unitedhealthgroup.com/newsroom/2019/2019-09-26-consumer-sentiment-survey-tech.html
[48] Longoni, C., et al. (2019). "Resistance to medical artificial intelligence." *JCR*. Retrieved from https://academic.oup.com/jcr/article-abstract/46/4/629/5485292?redirectedFrom=fulltext
[49] Longoni, C., & Morewedge, C.K. (2019). "AI can outperform doctors, so why don't patients trust it?" *HBR*. Retrieved from https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it
[50] Tran, V., et al. (2019). "Patients' views of wearable devices and AI in healthcare: findings from ComPaRe e-hort." *Npj Digital Medicine.* Retrieved from https://www.nature.com/articles/s41746-019-0132-y?utm_source=STAT+Newsletters&utm_campaign=f5d8c45344-health_tech_COPY_01&utm_medium=email&utm_term=0_8cab1d7961-f5d8c45344-149547853
[51] Balaram, B., et al. (2019). "Artificial intelligence: real public engagement." *RSA.* Retrieved from https://www.thersa.org/globalassets/pdfs/reports/rsa_artificial-intelligence---real-public-engagement.pdf
[52] KPMG. (2020). "Living in an AI world." Retrieved from https://advisory.kpmg.us/content/dam/advisory/en/pdfs/2020/living-in-ai-world.pdf

[53] Insights Team. (2019). "AI and healthcare: a giant opportunity." *Forbes.* Retrieved from https://www.forbes.com/sites/insights-intelai/2019/02/11/ai-and-healthcare-a-giant-opportunity/#2ea1e3be4c68

[54] Accenture. (2017). "Artificial intelligence: healthcare's new nervous system." Retrieved from https://www.accenture.com/t20171215T032059Z__w__/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf