

Excerpted and Summarized from:

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

Full white paper to be released in Spring 2020

### 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.<sup>1</sup> But this evolving field is complex, and as with all technologies, not without risk. As such, there is an urgent need to better communicate to clinicians, health system operators, and others information about how to harness the benefits of AI and mitigate the risks.

Regulators, purchasers, and users of traditional medical devices are accustomed to receiving specific information on design, materials, mechanism of action,\* and performance as they make decisions on approving, procuring, or using these technologies. However, for an array of logistical, technical, and legal/competitive reasons, manufacturers of AI-enabled tools may not be prepared or able to disclose some of this information.

This background document is meant to guide participants in the January 23<sup>rd</sup>, 2020 stakeholder meeting as they contemplate how AI-enabled clinical decision software differs from traditional medical products, and how these differences may affect supply and demand for information throughout product lifecycles. Discussion at the meeting will explore whether there are other ways to meet information needs when certain information can't be shared, and how best to incentivize development of safe and effective software products through clear communication about how they work.

### Background: What Makes AI-Enabled Clinical Decision Software in Health Care Different from Other Medical Products?

AI is a broad term that refers to the ability of a machine to perform a task that is normally done by humans. AI-enabled software can be designed in different ways, broadly classified into two categories. Rules-based algorithms use expert-derived rules, and defined and logical processes, to turn 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 so-called “labels” – for example, thousands of mammograms labeled with whether the patient was eventually diagnosed with cancer or not. The derived relationships can then be used to predict how new input data is likely to be labeled. Data-based AI is often referred to as “machine learning.”

---

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

Clinical decision software<sup>†</sup> that assists health care providers in making diagnoses, treatment decisions, managing population health, and carrying out general administrative duties have utilized rules-based AI for years. However, recent advances in machine learning are demonstrating the potential to significantly improve the performance of software in guiding more complex decision making and other tasks, thus opening the door to a range of new AI-enabled software.

However, software in general, and AI-enabled software in particular, differs from more traditional medical devices in critical ways that create challenges for regulators, as well as for clinicians, health systems and others who may wish to adopt the technologies. For example, some AI-enabled software products may produce clinical recommendations, but may 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.<sup>2</sup> Lack of trust may be exacerbated by concerns about where tort liability may fall if the software recommendation is wrong. Trade secrecy concerns may also limit the amount of information that companies that develop software are willing to disclose, both about how these systems work and how they were built.

Below, this paper describes three key differences between software and other medical products: (1) software uses 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 may lack an explanation of “how it works.”

### Health Data

Traditional medical devices act directly on the structure (or a function) of the body to produce results (although not through chemical action, which distinguishes devices from drugs). By contrast, software acts on health data, such as that produced through medical imaging, medical sensors such as electrocardiograms, or manually entered in electronic health records or other applications. These data can be incomplete, inaccurate, or biased.<sup>3</sup> For example, information gathered from electronic health records and fed into software systems 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.

Rules-based software often uses more limited, structured data for input, and data used in such software is generally more consistent. In contrast, data-based software often uses large, complex data sets as inputs; such data are more likely to have idiosyncrasies specific to particular work flows and individual physician, such as free text fields in EHR records. Patients’ access to care, including tests and procedures, will also affect the amount and types of health data available. Because this data is analyzed by software to reach recommendations, clear definitions around the data input requirements are necessary.

Data-based software, of course, also uses health data in the development of the system itself. Due to the heterogeneity discussed above, software that is developed with data from one location may not work at other locations without significant changes to the software program. Bias can also be a concern. If a software system isn’t trained with a representative dataset that contains sufficient numbers of patients from ethnic minorities or patients with co-morbidities, or if the recorded data is historically biased because

---

<sup>†</sup> 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.

of socioeconomic status, race, or other criteria, the resulting software may not work well with those populations and may even perpetuate existing biases within the health system.

A recent study found, for example, that an algorithm being used to predict patients' level of risk of serious illness used cost as a proxy for risk during training. Controlling for illness, white patients use the health care system more than black patients. The algorithm therefore assigned white patients higher risk scores than black patients who were equally ill, thereby reducing the number of black patients identified for extra care by more than half.<sup>4</sup>

Health data also rapidly changes format and terminology over time, and as new clinical practices and medical products come into use. As a result, performance of this software can degrade over time if it is not updated to account for these changes. As such, manufacturers and health systems will need to work together to monitor system performance and update software as needed.

### Rapid Software Development Cycles

Rapid updating makes software in general, and machine learning technologies in particular, distinctive among medical devices.<sup>5</sup> Manufacturers are able to act quickly to improve performance and correct problems found through real-world feedback, and then push updates to the users of those technologies. Certain types of machine learning software even have the potential to continuously update themselves in real-time, although clinical decision software of this type has not yet been approved or cleared by FDA. These updates are critical to not just improving the product, but also simply to maintain performance.

However, the rapid development cycle of software products is a challenge for regulatory agencies, which have set up review and clearance processes based on the slower development cycles of more traditional devices. In response, FDA has proposed 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.<sup>6,7</sup> 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,<sup>8</sup> and asked for input from the public regarding how to meet the challenges in regulating the AI-enabled software.

Frequent product updates also present concerns for adopters and users of these devices. Best practices still need to be developed to clearly inform software users of how updates may affect the safe and effective use of the product. In addition, global updates may affect local performance in unexpected ways, emphasizing the need for regular performance monitoring.

### Explainability

In the biopharmaceutical arena, there are products whose "mechanism of action" and/or "mode of action" are not known.<sup>9,10</sup> This sort of uncertainty is less common with traditional medical devices, though examples exist.<sup>11</sup> Some AI-enabled software products, however, may take uncertainty – and its attendant risks – to a higher level still.

Rules-based software is built on clearly defined, clinically-accepted rules and guidelines, although these guidelines may be based on observed statistical regularities rather than clear physiological mechanisms of action. However, 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 (often referred to as "black box" software).<sup>12</sup>

All medical products, including software, may fail in unusual, unpredictable ways when the mechanism of action is not clearly understood. In software, failures may be partly due to the fact that there are unforeseen patterns or “clues” present in the training data, which can produce 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 probably used site-specific features in the images to significantly influence the resulting prediction, rather than simply the underlying pathology. Because of these site-specific influences, the algorithmic models were not consistently generalizable to new health systems.<sup>13</sup>

When an explanation of how the software works cannot be provided, performance and risk must be carefully measured and understood. Prospective testing within the planned workflow is necessary to understand real-world performance and how the system may fail in unexpected ways. In addition, information regarding the certainty of a particular result, or what factors were weighed most heavily, may be helpful for users in understanding when to trust a particular result in the absence of a true explanation.

### Spectrum of Information on AI-Enabled Clinical Decision Software

There is a spectrum of information that various stakeholders may want around AI-enabled clinical decision software products: how a given 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 1).

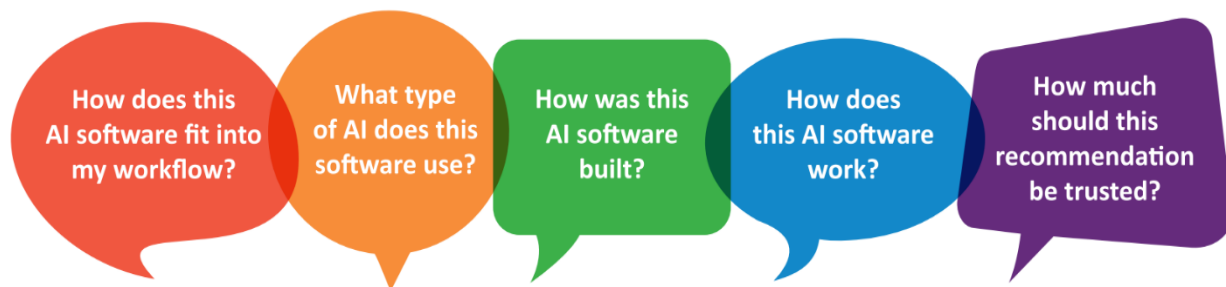


Figure 1 – Spectrum of information on AI-Enabled Clinical Decision Software. Various stakeholders throughout the total product lifecycle of a software product will want specific information of what the software does and how it fits into the workflow, what type of AI is used and how it was built, as well as information about how it works and when to trust the results.

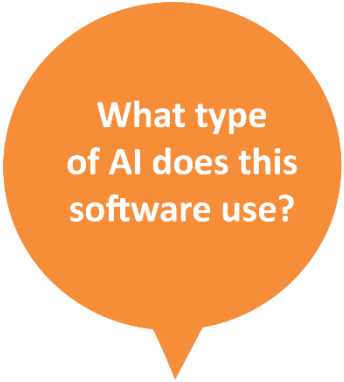
**How does this AI software fit into my workflow?**

Information about intended use should always be disclosed to all stakeholders. This baseline understanding should include the intended purpose of the device, how it is meant to be used and to fit into the workflow, and the intended significance of the results to clinical decision-making.

For example, all stakeholders will need to understand whether a given software product is designed to assist a clinician in decision-making or to automate that decision-making. If the software simply gives a doctor a notification 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, automated, or autonomous, software products will diagnose or treat patients directly. This automatic action may occur through hardware that is part of the system, such as an implantable cardiac


defibrillator analyzing heart rhythm and sending 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 these two categories may not be clear cut, as there are multiple gradations in between.



**What type of AI does this software use?**

In order to evaluate the software product properly, stakeholders also need to know what type of AI is being used in it. 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.<sup>14</sup>

Additionally, stakeholders will need to know if software developed with machine learning is locked, meaning that while the development used data-based techniques, the software does not continuously learn and change over time. As discussed above, we are not aware of any continuously learning standalone software products that have 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.



**How was this AI software built?**

Certain stakeholders may also want more detailed information about how the software was developed. 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.

Patents should, at least in theory, provide intellectual property protection even in the case of such full transparency.<sup>‡</sup> 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. Additionally, recent U.S. Supreme Court patent eligibility cases have made patenting of both medical diagnostics and software more difficult. When companies don't have secure patent protection, they may rely even more vigorously on trade secrecy to protect their competitive advantage.

In addition, neither patents nor copyright extends to raw data, which means that training data is commonly considered a trade secret. There also may be privacy implications in disclosing training data if it contains personal health information. If manufacturers are reluctant to disclose training datasets, summary information on the patient populations, including such information as demographics, social determinants of health, geographical region, comorbidities, and genetic markers, will still be useful. Having more of this

---

<sup>‡</sup> 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.

information may shed light on potential biases that will need to be tested for, and on whether the training population resembles the patient population of interest to the stakeholder.



**How does  
this AI software  
work?**


The most common question that stakeholders will have about AI-enabled clinical decision software is this one: how does the product work? Unfortunately, the information discussed above that would be required to reproduce the algorithms that drive a software product may not be helpful for human understanding of what that software is doing.

As discussed earlier, computer scientists use the term “explainability” (sometimes also referred to as “interpretability”) to convey whether a human-comprehensible understanding of an AI-enabled clinical decision software’s decision-making process exists. A true explanation delineates exactly how the software product will process input data into a result. Software that utilizes rules-based AI can always give “true” explanations, and certain types of machine learning can also be explained. For black box algorithms, statistical techniques that can produce a “likely” explanation are being explored.<sup>15</sup>

As noted earlier, lack of explainability has some precedent in medicine. In those situations, detailed performance data generally becomes more important to stakeholders. Indeed, robust performance data *should* be required by all stakeholders, regardless of the type of AI used, although requirements on rigor may differ based on risk. It is therefore critical for stakeholders to clearly communicate what type of performance data is being asked for and given. For example, when discussing a study involving data-based AI, it is important to understand whether the performance results are coming 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.<sup>16</sup> The latter will help test whether the software is depending on data features or patterns specific to the training dataset.

Stakeholders will also need to understand the methodology of any study to which the software was subjected, such as whether the study was retrospective or prospective, and whether the product was tested in the workflow with which it is intended to be used.<sup>17</sup> 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.<sup>18</sup>

Adopters will also need detailed information on how inputs into the software will need to 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. Stakeholders may also request to test the software on their data to evaluate local performance.



**How much  
should this  
recommendation  
be trusted?**

Finally, clinicians may desire specific information at the point-of-use about the results produced by a software system, to determine how heavily to weigh the results in their decision-making. This information could include the certainty of the software for a specific result, or the key input features that led to a specific recommendation. Users may also find it useful to have information about how many patients in the training data were similar to the patient or patients for whom this recommendation is being made. However, it is important that software systems be designed to communicate such information quickly, and in readily understandable ways, to accommodate clinicians' busy schedules.

## **Next Steps**

As noted above, stakeholders have always required substantial information about traditional medical devices to understand the full spectrum of risks and benefits, as well as to know how best to adopt and use them appropriately to advance patients' health. Some of the information needed by stakeholders to evaluate AI-enabled software is very similar to information required of other medical products, and manufacturers should be willing and able to provide it.

Yet, in other respects, AI-enabled software may raise these information demands to an entirely new level of detail and complexity, which could present technical challenges, both in conveying the information and making it understandable to human users. In addition, some of these stakeholder information needs or demands may also come into conflict with business and legal considerations confronting manufacturers.

In the final analysis, stakeholders/users of AI-enabled software will need to decide what new information they require to oversee and regulate these products and to use them safely and effectively. The January 23<sup>rd</sup>, 2020 forum will provide an opportunity for further discussion of the many issues described in this background paper, with an eye to laying out recommendations in a forthcoming white paper.



## References

---

- <sup>1</sup> 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>
- <sup>2</sup> 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>
- <sup>3</sup> 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](https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf)
- <sup>4</sup> 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>
- <sup>5</sup> 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>
- <sup>6</sup> FDA. (2017). "Digital Health Innovation Action Plan." *U.S. Department of Health and Human Services*. Retrieved from <https://www.fda.gov/media/106331/download>
- <sup>7</sup> 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>
- <sup>8</sup> 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>
- <sup>9</sup> Jilani, T.N., & Sharma, S. (2019). "Trihexyphenidyl." *StatPearls*. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK519488/>
- <sup>10</sup> Rosenbaum, S.B., & Palacios, J.L. (2019). "Ketamine." *StatPearls*. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK470357/>
- <sup>11</sup> 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>
- <sup>12</sup> 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>
- <sup>13</sup> 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>
- <sup>14</sup> 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>
- <sup>15</sup> Turek, M. (2018). "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency*. Retrieved from <https://www.darpa.mil/program/explainable-artificial-intelligence>
- <sup>16</sup> 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>
- <sup>17</sup> 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>
- <sup>18</sup> 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](https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf)