NEJM Catalyst | Innovations in Care Delivery

ARTICLE

AI-Enabled Clinical Decision Support Software: A "Trust and Value Checklist" for Clinicians

Christina Silcox, PhD, Susan Dentzer, David W. Bates, MD, MSc

Vol. 1 No. 6 | November — December 2020
DOI: 10.1056/CAT.20.0212

Machine learning and other forms of artificial intelligence (AI) are playing an increasing role in health care, particularly as an addition to human judgment in the form of clinical decision support (CDS). But as with all technologies, machine learning and AI will also have unintended consequences that could disrupt care and pose considerable risks for patients. It is vitally important that clinicians understand what is behind the recommendations that a CDS system offers and that any such system adds real value and enables clinicians to perform more effectively and efficiently in serving the needs of patients. This article presents a "trust and value checklist" that is aimed not at senior health system leadership, but rather at the clinicians who will be using these systems. The questions that the checklist poses include both those that the clinicians should be considering themselves and some that they will want to make sure that their leadership has addressed when making system selections. All of these questions should be considered, and answered to clinicians' satisfaction, before they start using and relying on CDS.

Introduction

Machine learning and other forms of artificial intelligence (AI) are playing an increasing role in health care, particularly as an addition to human judgment in the form of clinical decision support (CDS). Sophisticated software systems can now help spot early signs of disease, help interpret images, speed up diagnoses and improve their accuracy, and suggest appropriate treatment. It is axiomatic that these systems have some advantages over even the most proficient humans, as they can absorb vast amounts of information, detect important patterns, flag key anomalies, and advance tentative conclusions, all without becoming fatigued. But as with all technologies, machine

learning and AI will also have unintended consequences that could disrupt care and pose considerable risks for patients.[1]

Thus, clinicians should no more accept at face value the CDS they might receive from machine-learning systems than they would automatically accept the analysis or counsel of a clinical colleague. It is vitally important to be able to trust the source and to understand what is behind recommendations that the system offers, just as it is when obtaining a clinical peer's advice. It is also important that any CDS system add real value and enable clinicians to perform more effectively and efficiently in serving the needs of patients.

Choices about the selection of machine-learning systems will inevitably be made at multiple levels of a health care organization, including by its senior information technology leadership. This article presents a "trust and value checklist" that is aimed not at senior leadership but rather at the clinicians who will be using the system. The questions that the checklist poses include both those that they should be considering themselves and some that they will want to make sure that their leadership has addressed when making selection choices. All of these questions should be considered, and answered to clinicians' satisfaction, before they start using and relying on these systems. (See Table 1 for definitions of key terms used in this analysis.)

Overview of Machine-Learning/AI Applications in CDS

AI — the ability of a machine to perform a mental task normally done by humans — has been used in health care for years, mostly in the form of so-called "rules-based" algorithms and software. These systems use rules derived by experts, such as clinical guidelines, to turn "inputs" into an "output" — for example, drawing on electronic health record or other digitized information about patients to produce recommendations that they be screened for certain conditions.

> *"Sophisticated software systems can now help spot early signs of disease, help interpret images, speed up diagnoses and improve their accuracy, and suggest appropriate treatment."*

By contrast, newer forms of AI in health care today often incorporate "data-based" algorithms, known as machine learning. Machine-learning methods allow computers to "learn" patterns from a sample of data known as "training" data and then use the resulting mathematical operations underlying those patterns to process new data to make predictions.

Most machine learning used in health care is based on so-called "supervised" learning, in which the algorithm initially learns on a set of data, known as training data, that links inputs to outputs. For example, an input, such as a mammogram image, is linked to an output, such as the presence or absence of a malignant tumor. These training data are "labeled," in that each input in the training data set (mammogram A) is linked to a label of the output (cancerous tumor type X, previously established as being present in that image). The learning algorithm is exposed to large numbers of inputs linked to various outputs. The software then calculates statistical relationships between the

Table 1. Glossary of Terms

Term	Definition
Artificial intelligence (AI)	A broad term meaning a machine's ability to perform a task normally done by humans.
Machine learning	A subset of AI, machine-learning algorithms build a mathematical model based on sample data, known as "training data," to make predictions or recommendations without being explicitly programmed to do so.
CDS software	Software that is intended to provide decision support to clinicians for the diagnosis, treatment, prevention, cure, or mitigation of diseases or other conditions.
"Black box"	Software that does not explain how input data are analyzed to come to a recommendation; this lack of explanation may be because the algorithm or model is too complex to be understood by humans, or because the functionality is considered proprietary.
Explainability	A human-comprehensible explanation of how a model combines and uses data inputs to come to a specific recommendation; also referred to by computer scientists as model "transparency."
Labeled Data	Training data that are paired with what the software developer considers the correct "output" for each input example.
Supervised learning	A type of machine learning that uses labeled data sets to create an algorithm that can predict the "label" when new data are entered.
Locked model	A function or model that was developed through data-based AI methods but does not update itself in real time; differs from "unlocked" or "continuously learning" algorithms that may be developed in the future and that may be able to adjust better to changes in data and differences in local conditions.
Software as a medical device	Software intended to be used for one or more medical purposes and to perform these without being integral to the hardware of a medical device; differs from "software in a medical device," in which software is integral to the hardware of a medical device.

Source: The authors.

input data and the associated label, allowing it to predict how new input data would be labeled — for example, that a given new mammogram, fed into the system, shows signs of cancer.

Software developers can choose different ways to label or "annotate" outputs. For example, a software developer may label an image as "malignant" if a set of three board-certified radiologists agree that one X-ray contains evidence of a tumor, or the developer may rely on the results of biopsies to determine whether to label images as malignant or not. This process of clinical annotation, or labeling, is typically the most costly and time-consuming aspect of developing a machine-learning software system.

Once the algorithm is trained on a data set, its performance is tested by its developer, typically first by subjecting the software to a subset of the original data set that has been collected for training purposes but that is not actually used during the training process. For software that is considered "low risk" — for example, not likely to be used directly in clinical decision-making in a way that could adversely affect a patient's health, such as predicting a patient's risk of a hospital readmission — this degree of performance testing might be sufficient. For higher-risk systems that could be used in clinical decision-making that directly affects patients' health — for example, predicting a hospitalized patient's risk of sepsis — algorithms should be tested on independent data sets not collected at the same site, or at the same time, at which the original training data were collected. In the sepsis example, such testing could require using data from a different health system with a very different patient population to ensure that the software was able to make appropriate recommendations in different circumstances.

Most of the machine-learning–enabled clinical decision software on the market today is "locked" software, in which the algorithm, having been trained on an initial data set and tested, does not continuously learn and automatically adapt thereafter (although these products may be frequently

updated by the developer). To date, in fact, this is the only type of "software as a medical device" that the FDA has allowed onto the market. The future may bring more "unlocked" or "continuously learning" algorithms that may be able to adjust better to changes in data and local conditions than locked software systems; users will then have to be aware that the output of the software could change over time, even given the same input data, because of these self-adjustments.[2] The FDA is still considering how to regulate these types of algorithms.[3]

Machine-Learning Systems on the Market Today

With exceptions like the sepsis algorithm cited above, the top machine-learning systems on the market today typically involve imaging or pathology, because systems can be readily trained to identify visual differences among large numbers of scans or slides. One example of a machine-learning system now on the market is IDx-DR, which was built to detect signs of diabetic retinopathy and was granted de novo status by the FDA[4] in 2018, permitting the software to be used in primary care settings. The device's software analyzes retinal images taken by a built-in camera for signs of hemorrhages, microaneurysms, or other indications of the condition. The initial learning algorithm was trained by exposing it to thousands of retinal images, which allowed it to "learn" how to spot the appropriate signs.

> *"Clinicians should no more accept at face value the CDS they might receive from machine-learning systems than they would automatically accept the analysis or counsel of a clinical colleague."*

In the clinical trial that preceded the FDA's clearance of IDx-DR, the device demonstrated 87% sensitivity and 91% specificity at detecting more than mild diabetic retinopathy in retinal images.[5] Within a minute after snapping the retinal image, the device's operator — who may be a nurse, physician assistant, or someone else — receives notification that the image is too low quality and must be retaken; that the patient should be retested in 12 months; or that the patient has more than mild retinopathy and should be referred to an ophthalmologist or other eye care professional for follow-up care.

A second example of machine learning in health care is software developed by a company called Viz.ai, Inc., which links to computer tomography scanners and continuously analyzes brain images of patients for early detection of strokes. The software alerts specialists within several minutes that a suspected stroke has been identified and sends the images to their smartphones for viewing and initial treatment decisionmaking. The software has been cleared for clinical use by the FDA for detection of large vessel occlusions or blockages in carotid or cerebral arteries that cause up to 40% of ischemic strokes; for automated analysis of cerebral perfusion images; and for detection of suspected intracranial hemorrhage (ICH). A retrospective study conducted by the company of the ICH application showed sensitivity of 93% and specificity of 90%. The software recently became the first AI application to be deemed eligible by the Centers for Medicare and Medicaid Services for a New Technology Add-on Payment, which enables reimbursement of health systems that employ it in treatment of Medicare patients by up to $1,040 per use.

Clearly, these technologies can be extraordinarily valuable for clinicians and patients, but they also raise both obvious and not-so-obvious concerns.

For one thing, once the clinical decision software is trained and locked, it requires that new health data be entered into it so that the software can make its recommendations. But in varying circumstances, that input data may be incomplete, inaccurate, biased, out of date, not structured, or not defined in a way that the system is expecting — in other words, the data will reflect the messy "real world" of health care. For example, if a treatment recommendation system is trained partly with diagnosis data coded in *International Classification of Diseases* (ICD), 9th revision, but is fed input data coded in ICD-10 instead, the system may produce treatment recommendations that are incorrect. Unfortunately, these types of data disconnects are all too common in health care.[6]

For another, it is often unclear what information these systems are using to form the basis of their decisions. A system trained on data from a single institution may, for many reasons, not perform as well when presented with data from other institutions.

For example, when a software system at Mount Sinai Medical Center in New York was built to recognize pneumonia in chest X-rays, the system was effectively informed by such anomalies as a high rate of pneumonia prevalence among inpatients and widespread use of portable X-ray machines for patients considered too sick to get out of bed. (Portable X-rays produce somewhat different-looking images from conventional X-rays.) When the algorithm was tested on a different set of X-rays, but ones still from Mount Sinai, it accurately detected pneumonia 93% of the time. But when it was tested on data from other sites, with far lower rates of pneumonia and far less use of the portable equipment, the success rate fell to just 73% to 80%.[7] Only in retrospect did it become clear that the unique circumstances at Mount Sinai had in effect been confounding factors.

Another related issue is the "explainability" of machine-learning software — that is, whether its developers, or even the system itself, can provide a comprehensible explanation of how the software weighs and combines inputs to come to a result. Most, although not all, machine-learning systems are "black boxes" — that is, there is no human-comprehensible explanation for how the software actually functions and turns inputs into outputs.[2] Although clinicians often do not understand fully how a piece of medical equipment works — for example, how an MRI creates an image — they generally have the assurance that someone in their health system does understand the inner workings and can vouch for the reliability of the equipment. But with black box software systems, even the developers will not be able to explain how they work — somewhat similar to instances in which pharmaceutical drugs lack a fully understood mechanism of action.

> *"For higher-risk systems that could be used in clinical decision-making that directly affects patients' health — for example, predicting a hospitalized patient's risk of sepsis — algorithms should be tested on independent data sets not collected at the same site, or at the same time, at which the original training data were collected."*

Beyond the reliability or explainability of these software systems lies an even larger issue that pertains to the evolving health data ecosystem, albeit not one within any given clinician's, or even health system's, control. There is clearly exponential growth in the collection of health and health care data, which may be initially housed in electronic health records under strict privacy laws and protocols or collected through applications operating outside the Health Insurance Portability and Accountability Act of 1996 or other privacy statutes.

Increasingly, these types of data are transferred and stored in the "cloud" — collections of data centers available to many users over the Internet and generally owned by large technology companies. There, data may be processed with the aid of cloud computing tools and techniques that are proprietary to these large companies. It is simply unknown how private the data really are and how secure from any possible manipulation or data breaches, either at the individual electronic health record level or in the cloud.

There is, of course, very little that any single clinician or health system can do about this issue and the uncertainties it highlights. But it is worth bearing in mind, as the world of machine learning and AI evolves, that the provenance of the data that these systems are based on may continue to pose a concern.

Are CDS Systems Trustworthy?

Considering these complex considerations, can CDS software enabled by machine learning be trusted? Is there a "*Good Housekeeping* Seal of Approval" or Underwriters Laboratories–type label to warrant that these systems are sound?

One might suppose that some federal regulatory agency such as the FDA should regulate all CDS software, but for better or worse, no agency does that. By law, in fact, the FDA cannot regulate certain types of CDS systems: specifically, those that are designed to allow clinicians to review independently any recommendations that the system offers, so that they are still relying primarily on their own clinical judgment. The 21st Century Cures Act codified into law that much of this type of software, as well as other software unrelated to prevention, diagnosis, or treatment of disease, is not considered a medical device and is not subject to the FDA's regulatory authority.[8] An example might be software that helps health systems determine how to move patients more efficiently through a hospital.

By contrast, the FDA can regulate — in that it "clears" or "approves" — so-called "software as a medical device." These are systems that are intended to "support or provide recommendations" to health care providers about prevention, diagnosis, or treatment of medical conditions. Even in this respect, however, the FDA operates in a vast gray area, and it has used a relatively light hand on regulation to date. For some software that it deems "low risk," in fact, the FDA has chosen under its enforcement discretion authority not to enforce compliance with medical device regulatory requirements that could otherwise compel active FDA scrutiny and clearance. An example that the FDA has cited as low risk is a machine-learning algorithm that trends and classifies patients' data, such as blood test results and weight, to assist a clinician in flagging cholesterol-management issues for patients.

> *Input data may be incomplete, inaccurate, biased, out of date, not structured, or not defined in a way that the system is expecting."*

In 2019, to clarify its regulatory role and expectations of software developers, the agency issued draft guidance setting forth its proposed regulatory approach to these software systems. It said that software developers are under an obligation to provide clarity and transparency about how their systems operate. It asserted that "the software developer should describe the underlying data used to develop the algorithm and should include plain language descriptions of the logic or rationale used by an algorithm to render a recommendation." The developer should also identify the "sources underlying the basis for the recommendation," such as clinical practice guidelines.

The FDA has asked for comments on this draft guidance and is not likely to issue final guidance for some time. In the meantime, although it continues to clear or approve CDS systems, it is not completely clear which systems need a green light from the FDA to come onto the market. Thus, it is not an ironclad rule that a software developer has to obtain FDA clearance to market an AI product. Conversely, although FDA clearance or approval provides substantial comfort that a system is sound, it is not a guarantee against errors and adverse events resulting from these systems.

The Trust and Value Checklist

Amid these many uncertainties about CDS software, it thus seems reasonable to create a checklist of questions. This checklist can serve a dual purpose. It includes questions that careful clinicians should ask in order to decide whether a system is trustworthy enough that they might even rely on it if its recommendations conflicted with their initial clinical thinking. It also includes questions to be explored by health system leaders charged with making purchase decisions about the software.

We term this a trust and value checklist because we believe that both components of this equation are important. Although software meant to inform and augment clinical decision-making should not be followed blindly, if the software never changes a user's mind, then it is not adding value. A CDS system worth purchasing will at least occasionally clear a high bar: "I'll trust your recommendations even if I initially thought otherwise," as one might say to a highly respected clinical colleague.

Although some of these questions speak directly to underlying machine-learning technology, others deal with issues such as usability or the ethics underpinning the collection of data that informs these systems. We believe that clinicians and health system leaders will feel reasonably comfortable that they understand the risks and the benefits of adopting a system if they can answer the following 10 questions to their own satisfaction. We have divided the questions into several categories.

The Value of the CDS System

1. Does use of this AI-enabled CDS software or system yield clinically important improvement compared to the status quo?

This question should be the first one asked of any medical product. For AI-enabled software, "improvement" could address any of the domains of the Quadruple Aim: improving individual or

population health, enhancing the experience of care, reducing costs, and improving providers' work life.[9] Software that simply automates certain tasks may still enable clinicians to spend more time with patients and focus on other aspects of care, granting them what Dr. Eric Topol of the Scripps Research Translational Institute refers to as the "gift of time."[10] A threshold question should be whether a new software product or system truly meets one or more of these goals.

2. Will use of this AI-enabled CDS software fit into my clinical workflow and/or my team's workflow?

Tools are valuable only if they are used, which is unlikely if their use is awkward, inconvenient, overly time-consuming, or taxing in any other way.[11] The history of adoption of electronic health records underscores that clinicians will resent any technology that does not fit into their existing clinical work processes, or workflow — and may only be willing to change their own modes of practice to adopt a new tool or technology if there is a very large clinical benefit. Thus, threshold questions for CDS software should include the following:

- Does this tool make sense within my current workflow?
- Does it make recommendations in a timely manner?
- Does it require me to manually enter information, move to another screen, or otherwise make additional "clicks"?
- If so, does the clinical benefit warrant this additional activity and possible annoyance?

3. What information does the AI-enabled CDS software provide at the point of use about the logic behind the decisions or recommendations that it produces and the degree of certainty that these recommendations are correct?

As is noted above, clinicians need to be able to trust the software they are using, as well as understand when some results may be less reliable for certain patients. It is also important to know what information the software may provide in addition to the result so that users can understand how to interpret the result. For example, the software could indicate how certain it is about a specific recommendation or specify key input elements that influenced the recommendation. A CDS software system evaluating mammograms might not only identify areas in the image that could be cancerous but also calculate and tag them with a score indicating how probable it is that the areas are cancerous. If systems operate this way, all of the information should be displayed in a way that users can quickly digest and interpret, such as differentiating problematic areas by color.

> *"With black box software systems, even the developers will not be able to explain how they work — somewhat similar to instances in which pharmaceutical drugs lack a fully understood mechanism of action."*

At the same time, the information that clinicians will need about the certainty of recommendations from CDS systems is likely to differ on the basis of the risks of relying on the system. Software that

can influence treatment decisions — such as recommending a specific cancer treatment protocol or thrombolytic therapy for stroke — requires more extensive explanation and information to gain clinician "buy-in" than does software that automates surgical room scheduling, for example. This additional information around the software results will be particularly important when the physician will be taking on medical liability for a decision.[12] Ideally, software should supply a confidence interval around the likelihood that any individual prediction is correct. Physicians will probably want to regard recommendations with more suspicion when the confidence intervals around them are broad (e.g., 0.35 to 0.99).

The Data Behind the CDS System

4. What was the source of the data used to develop and train the AI-enabled CDS software?

As is suggested above in the Mount Sinai example, it is important for clinicians to understand what type of data a machine-learning system has been trained on and how similar those data are to the health system in which a clinician is working. Many current AI-enabled CDS systems use imaging data, but the machines are not simply learning from the images themselves. All sorts of other factors — for example, what types of patients tend to have their images taken, what equipment is used, what positions the patients are placed in, and what times the scans are done — may all create site-specific patterns that the learning algorithm detects and uses in its recommendations. As a result, training data from one site may be skewed in such a way that they lead a system to make inappropriate recommendations based on data from another site.

5. How were the training data labeled?

As is noted above, in machine-learning systems that are "supervised," the learning algorithm is exposed to large numbers of inputs — for example, mammogram images that are associated with a label that represents a particular output such as a malignant tumor. The ways in which data are labeled are critical to an algorithm's performance, and understanding how is central to guarding against inadvertent bias in the training data.

For example, when one element of a software suite called Impact Pro, a software program produced by Optum, was used by one health system to predict which patients with complex medical needs could benefit from additional supports, it systematically overestimated the health status of African American patients.[13] The reason: this part of the software used health costs as a proxy for health needs. But because African Americans suffer from structural disparities in the health care system and are often undertreated, their costs in fact are lower, even though they may be very sick. Thus, few African Americans in effect would be identified as potential beneficiaries of the additional supports based on use of this single element of the software alone. Optum has said the software is not systematically biased and evaluates many variables other than costs, and that it discourages using Impact Pro selectively in the way the health system did.

Understanding how training data are labeled thus allows thinking through how labeling might bias recommendations. Variations in labeling also underscore the importance of testing software, as is described further below.

6. Does this AI-enabled CDS software appropriately respect patient privacy?

Patients' privacy could potentially be compromised in two main ways: first, from the standpoint of the data used to train and test the system, and second, from the standpoint of the data that clinicians may supply about their patients to enable the software to make a recommendation.

> *"By law, in fact, the FDA cannot regulate certain types of CDS systems: specifically, those that are designed to allow clinicians to review independently any recommendations that the system offers, so that they are still relying primarily on their own clinical judgment."*

If the system is trained on data traceable directly to specific patients, it will be important to ascertain that these data were obtained with patients' consent. Next, health systems purchasing software should know whether the clinical data that will be entered into it — either manually or through connections to electronic health records or other health data systems — to generate recommendations are also accessible to any other party, so that they can take steps to assure the confidentiality and privacy of data at that point. Health systems should also know whether these data are identifiable by patient or de-identified, whether the data have been appropriately secured, and for what purposes the data may be used. As was noted earlier, there are also larger issues about data privacy and security that are beyond the control or influence of individual clinicians.

Testing

7. Does this AI-enabled CDS software fall under the FDA's regulatory authority, and, if so, has it been cleared or approved by the agency?

As is noted above, some, but not all, CDS software falls under the FDA's regulatory aegis, and the FDA has enormous discretion about what and how it will regulate. Software that does fall under the FDA's purview generally requires some active degree of scrutiny before it is "cleared" or "approved" for entry onto the market. However, for some low-risk software, the FDA has chosen under its enforcement discretion capabilities not to enforce compliance with medical device requirements that would otherwise compel active FDA scrutiny and clearance. Clinicians should familiarize themselves with the testing that is required for FDA clearance or approval of a specific product, because the requirements will differ on the basis of the FDA's assessment of the product's risk.

8. If a software system is not under the FDA's authority, how was it tested, against what standard, and by whom (other than by its developers)?

Independent of the FDA or other regulatory authority review, it is important to understand how else AI-enabled CDS software has been tested, and against what standards, by entities other than the software developers. (These test results are reported statistically, often in terminology that clinicians may find foreign to them; a concept such as positive predictive value, e.g., may be referred to as a "c-statistic."[14]) Although developers will test a system's performance, other entities

— including health care systems considering adopting the system — will also want to test it (as described further below). What is more, because the language used to describe testing of AI algorithms is not yet standardized, clinicians should ask detailed questions about testing methodology to understand what the performance results actually mean.

It is particularly important to understand what data were used to generate the performance statistics about a software system. In some instances, these will be based on tests performed on a subset of the same data on which the algorithm was trained. (As is noted above, this data subset would have been identified and separated from the larger training data set before training began.) Alternatively, performance statistics can be derived on the basis of data from a completely independent data set, such as that from a different health system. Understanding which data sources were used to generate the performance data will help to determine whether the software system may be "overfitted" to the training data.

The above example about chest X-rays illustrates the point: algorithms sometimes learn from the detail and noise in the training data all too well, to the extent that the learning negatively affects the performance of the model when it encounters new data from a different health system or other context. That is the overfitting phenomenon. When such overfitting occurs, the algorithm does not perform as well when it is exposed to new data from another setting. Clinicians need to know that a software system performs well when it is exposed to new data, not just a subset of the data with which it was trained.

9. Has this AI-enabled CDS software been tested on data from my own hospital or health system?

This question is a threshold one for decision-makers in charge of purchasing software for a system, but clinicians within the system should ask it as well. Building on the discussion above about performance statistics, a given software product absolutely must be tested — either before it is acquired or during implementation — with data from the health care system in which it will be used. This process will not only settle the question of overfitting versus underfitting but will also give clinicians the best sense of how well the software works with data collected in their own work processes, with their specific technologies, and with their specific patient population.

> *"We believe that clinicians and health system leaders will feel reasonably comfortable that they understand the risks and the benefits of adopting a system if they can answer the following 10 questions to their own satisfaction."*

The most accurate test results will be achieved if this testing is carried out within the conditions that it will be used (e.g., as an assistive technology that augments clinical judgment). Clinicians and health care systems should use data collected from their own patients, in real time, and feed it into the software just as they would do in the course of practice. Only by doing so will they see and understand how use of the software and its algorithm will affect their clinical performance and do or do not provide clinical utility and value.[15] Clinicians and health systems should carefully evaluate

these results before adopting these AI-enabled software systems and also have plans in place to evaluate these systems again once they are implemented in their health system.

Maintaining and Improving the Software Over Time

10. What provisions has the developer made for monitoring and updating the AI-enabled CDS software over time to account for potential degradation in performance?

Users of smartphones and computers are used to regular and frequent software updates, but the regulatory regime for medical devices was not designed for the typical software product update cycle of weeks to months. The FDA is exploring new methods for regulating these products,[3] but in the meantime, clinicians and provider systems should ask developers how and how frequently the software will be updated — and how any new training data used to make these updates may differ from how the original training data were collected and labeled. Provider systems may also want to consider testing software performance regularly with newer data derived from their own patient population, to understand whether their data are changing in ways that cause the software performance to degrade.

AI and machine learning have the potential to revolutionize health and health care and to improve safety, quality, and value substantially. But clinicians and health systems will need to become much more knowledgeable about these technologies and will need to exercise a form of self-regulation in adopting and using them. The above trust and value checklist should assist with that process.

It may be unrealistic to expect that most technologies under consideration will "check all the boxes" or score perfectly. Moreover, it may be acceptable for software that poses fewer direct risks to direct patient care — for example, that bears on operational issues in health systems, as opposed to clinical decision-making — to have less than perfect scores. It is also unlikely that individual clinicians can carry out these assessments on their own. Provider systems should establish AI/machine-learning review committees that are similar in their general function and purpose to Pharmacy and Therapeutics committees, to provide assurance that advanced CDS software truly lives up to its much-vaunted promise.

Although we have termed these questions a trust and value checklist, there are other considerations in purchasing systems that we have not addressed, such as return on investment. Those calculations are more appropriately the domain of health systems leaders and of chief financial and clinical officers working together. As machine-learning–enabled CDS software comes into broader use, however, a working familiarity with the issues identified through this checklist will help clinicians become comfortable using these tools to practice in a rapidly transforming era of health care.

Christina Silcox, PhD
Managing Associate, Duke-Margolis Center for Health Policy, Duke University, Durham, North Carolina, USA

Susan Dentzer
Senior Policy Fellow, Duke-Margolis Center for Health Policy, Duke University, Durham, North Carolina, USA

David W. Bates, MD, MSc
Chief, Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, USA

Disclosures: Christina Silcox has nothing to disclose. Susan Dentzer discloses that she has hosted a 10-part podcast series for Optum that includes references to its rules-based predictive analytics platforms. David W. Bates discloses consulting fees from EarlySense and CDI-Negev, Ltd, as well as equity from Valera Health, CLEW Medical, MDClone, and AESOP Technology.

References

1. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc 2017;24:246-50 https://pubmed.ncbi.nlm.nih.gov/28011595/.

2. Consumer Technology Association. Definitions/characteristics of artificial intelligence in health care (ANSI/CTA-2089.1). February 2020. Accessed October 6, 2020. https://shop.cta.tech/products/definitions-characteristics-of-ai-in-health-care.

3. U.S. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device. 2019. Accessed October 6, 2020. https://www.fda.gov/media/122535/download.

4. U.S. Food and Drug Administration. De novo classification request. November 20, 2019. Accessed October 6, 2020. https://www.fda.gov/medical-devices/premarket-submissions/de-novo-classification-request.

5. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 https://pubmed.ncbi.nlm.nih.gov/31304320/ https://doi.org/10.1038/s41746-018-0040-6.

6. Wright A, Wright AP, Aaron S, Sittig DF. Smashing the strict hierarchy: three cases of clinical decision support malfunctions involving carvedilol. J Am Med Inform Assoc 2018;25:1552-5 https://doi.org/10.1093/jamia/ocy091.

7. Couzin-Frankel J. Artificial intelligence could revolutionize medical care. But don't trust it to read your x-ray just yet. Science. June 17, 2019. Accessed October 6, 2020. https://www.sciencemag.org/news/2019/06/artificial-intelligence-could-revolutionize-medical-care-don-t-trust-it-read-your-x-ray.

8. U.S. Food and Drug Administration. Changes to existing medical software policies resulting from Section 3060 of the 21st Century Cures Act. September 27, 2019. Accessed October 6, 2020. https://www.fda.gov/media/109622/download.

9. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12:573-6 https://pubmed.ncbi.nlm.nih.gov/25384822/ https://doi.org/10.1370/afm.1713.

10. Health Education England. Preparing the healthcare workforce to deliver the digital future. February 2019. Accessed October 6, 2020. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf.

11. Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10:523-30 https://pubmed.ncbi.nlm.nih.gov/12925543/ https://doi.org/10.1197/jamia.M1370.

12. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019; 322:1765-6 https://pubmed.ncbi.nlm.nih.gov/31584609/ https://doi.org/10.1001/jama.2019.15064.

13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53 https://pubmed.ncbi.nlm.nih.gov/31649194/ https://doi.org/10.1126/science.aax2342.

14. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322:1806-16 https://pubmed.ncbi.nlm.nih.gov/31714992/ https://doi.org/10.1001/jama.2019.16489.

15. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356:1399-409 https://pubmed.ncbi.nlm.nih.gov/17409321/ https://doi.org/10.1056/NEJMoa066099.