Check for updates

Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations

David B. Larson, MD, MBA[a], Hugh Harvey, MBBS[b], Daniel L. Rubin, MD, MS[c], Neville Irani, MD[d], Justin R. Tse, MD[e], Curtis P. Langlotz, MD, PhD[f]

Abstract

Although artificial intelligence (AI)-based algorithms for diagnosis hold promise for improving care, their safety and effectiveness must be ensured to facilitate wide adoption. Several recently proposed regulatory frameworks provide a solid foundation but do not address a number of issues that may prevent algorithms from being fully trusted. In this article, we review the major regulatory frameworks for software as a medical device applications, identify major gaps, and propose additional strategies to improve the development and evaluation of diagnostic AI algorithms. We identify the following major shortcomings of the current regulatory frameworks: (1) conflation of the diagnostic task with the diagnostic algorithm, (2) superficial treatment of the diagnostic task definition, (3) no mechanism to directly compare similar algorithms, (4) insufficient characterization of safety and performance elements, (5) lack of resources to assess performance at each installed site, and (6) inherent conflicts of interest. We recommend the following additional measures: (1) separate the diagnostic task from the algorithm, (2) define performance elements beyond accuracy, (3) divide the evaluation process into discrete steps, (4) encourage assessment by a third-party evaluator, (5) incorporate these elements into the manufacturers' development process. Specifically, we recommend four phases of development and evaluation, analogous to those that have been applied to pharmaceuticals and proposed for software applications, to help ensure world-class performance of all algorithms at all installed sites. In the coming years, we anticipate the emergence of a substantial body of research dedicated to ensuring the accuracy, reliability, and safety of the algorithms.

J Am Coll Radiol 2021;18:413-424. Copyright © 2020 American College of Radiology

[a]Vice Chair, Education and Clinical Operations, Department of Radiology, Stanford University School of Medicine, Stanford, California.
[b]Institute for Cognitive Neuroscience, University College, London, UK.
[c]Director of Biomedical Informatics at Stanford Cancer Institute, Departments of Biomedical Data Science, Radiology, and Medicine, Stanford University School of Medicine, Stanford, California.
[d]Department of Radiology, University of Kansas Medical Center, Kansas City, Kansas.
[e]Department of Radiological Sciences, David Geffen School of Medicine, University of California, Los Angeles, California.
[f]Associate Chair, Information Systems, Department of Radiology, Stanford University School of Medicine, Stanford, California.
Corresponding author and reprints: David B. Larson, MD, MBA, Stanford University School of Medicine, Department of Radiology, 300 Pasteur Dr, Stanford CA 94305-5105; e-mail: david.larson@stanford.edu.

Dr Larson reports research funding, Siemens Healthineers; patent, Cincinnati Children's Hospital; and Shareholder and Advisory Board: Bunker Hill Health. Dr Harvey reports salary from Hardian Limited, personal fees from Smart Reporting GmBH, nonfinancial support from Algomedica and Segmed.ai, and an honorary research fellowship at University College London, outside the submitted work. Dr Rubin reports grants from Philips, grants from GE, personal fees from Genentech, outside the submitted work. Dr Irani reports testing started for evaluating chest x-ray AI algorithm strengths and limitations. Dr Tse reports grants from GE Healthcare/RSNA Resident Research Grant, outside the submitted work. Dr Langlotz reports nonfinancial support from Nines.ai; personal fees and nonfinancial support from whiterabbit.ai; nonfinancial support from GalileoCDS, Inc; nonfinancial support from Bunker Hill, Inc; grants from GE Healthcare; grants from Philips Healthcare; grants from Siemens Healthineers; nonfinancial support from Sirona Medical; grants from Carestream; grants from IDEXX; grants from Google; grants from IBM; grants from Hospital Israelita Albert Einstein, outside the submitted work. The other authors state that they have no conflict of interest related to the material discussed in this article. Dr Larson, Dr Rubin, Dr Irani, Dr Tse, and Dr Langlotz are nonpartner, non–partnership track employees. Dr Harvey is a partner at Hardian Limited.

1546-1440/20/$36.00 ■ https://doi.org/10.1016/j.jacr.2020.09.060

INTRODUCTION

Artificial intelligence (AI) algorithms hold promise for improving care, especially in imaging diagnosis [1,2]. Robust evaluation of AI-based software before implementation is needed to reduce patient and health system risk, establish trust, and facilitate wide adoption. [3]. Regulators have proposed frameworks for ensuring the safety and effectiveness of AI-based software as a medical device (SaMD) [4-12]. These frameworks provide a solid regulatory foundation but also have shortcomings that are likely to limit adoption of these algorithms in practice.

In this article, we review the major regulatory frameworks for SaMD applications, identify major gaps, and propose additional strategies to improve the development and evaluation of diagnostic AI algorithms. This article represents the joint perspective of a small group of radiologists concerned about the safety, reliability, and sustainability of AI-based diagnostic algorithms in the clinical environment. The views we represent are our own and not necessarily those of the organizations in which we serve. However, we urge all stakeholders and the organizations they represent to align these recommendations with their current endeavors.

SUMMARY OF THE FDA, INTERNATIONAL MEDICAL DEVICE REGULATORS FORUM, AND EUROPEAN UNION FRAMEWORKS

Regulators of SaMD applications, including the FDA in the United States, have been guided by the Global Harmonization Task Force, established in 1993, and the International Medical Device Regulators Forum (IMDRF), which superseded the Global Harmonization Task Force in 2012. These voluntary groups of device regulators proposed key definitions [4], risk categories [5], a quality management system [6], and standards for clinical evaluation and investigation [7-9].

The IMDRF has defined SaMD as “software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device” [4]. This includes software that is intended for the medical purpose of “diagnosis, prevention, monitoring, treatment, or alleviation of disease,” among other purposes [4], including AI-based diagnostic algorithms for medical imaging.

The IMDRF has proposed that the degree of regulatory scrutiny of algorithms should be based on the risk for harm, recommending four risk categories for SaMD applications. These categories depend on the health care condition severity (urgent, serious, or critical) and information provided by the SaMD application to the health care decision (to inform clinical management, to drive clinical management, or to treat or diagnose) [5].

To ensure the safety, effectiveness, and performance of SaMD, the IMDRF has outlined quality management system principles for SaMD applications [6]. These include (1) an organizational structure that provides appropriate leadership, accountability, and governance; (2) a set of SaMD life cycle support processes, embedded in product planning, risk management, documentation, configuration, measurement, and outsourcing; and (3) a set of realization and use processes, including requirements management, design, development, verification and validation, deployment, maintenance, and decommissioning [6].

According to the IMDRF [7], clinical effectiveness evaluation should be accomplished in three stages: (1) establishment of a valid clinical association between the SaMD output and the targeted clinical condition, which can be based on existing clinical evidence or following generation of new evidence; (2) analytical validation, referring to the ability of the algorithm to reliably process input data to generate the intended output data; and (3) clinical validation, referring to the ability of the algorithm to have a meaningful clinical impact [7]. Manufacturers are expected to continuously improve the performance of the application throughout its life cycle using real-world performance data [7].

The IMDRF also recommends that a clinical evaluation report be compiled to outline (1) the technology on which the device is based; (2) the intended use of the medical device and any claims made about its safety, clinical performance, and effectiveness; and (3) a description of the clinical data and how it demonstrates the safety, clinical performance, and effectiveness of the device [7]. The European Union has codified this requirement for a clinical evaluation report and also requires manufacturers to prepare and follow a postmarket follow-up plan [10].

In January 2019, the FDA published as a working model of the concept of a software precertification program—a voluntary pathway for manufacturers of SaMDs who have demonstrated a robust culture of quality and organizational excellence and are committed to monitoring real-world performance [11]. Later in 2019, the FDA proposed a framework to streamline improvements in SaMD applications [12]. During the initial premarket review, manufacturers submit a “predetermined change control plan” containing two key elements: (1) SaMD specifications for the changes the manufacturer expects to make after deployment and (2) a step-by-step algorithm change protocol that delineates the data and the procedures necessary for postrelease testing and refinement [12].

GAPS IN THE CURRENT REGULATORY FRAMEWORKS

These regulatory frameworks address many key aspects to ensure safety, effectiveness, and performance of SaMD applications, especially focused on manufacturers' responsibilities. However, a number of gaps remain.

Conflation of the Diagnostic Task and the Diagnostic Algorithm

Although an algorithm and the task it performs are closely linked, they are separate entities. For example, an algorithm that automatically classifies renal cysts identified on CT according to the Bosniak classification system [13] is not, itself, the Bosniak classification system; rather, it is the vehicle for applying the classification system to the image.

This distinction between task and algorithm also applies to clinical assessments (such as the presence or absence of pneumonia on a chest radiograph) and measurement systems (such as a measure of left ventricular ejection fraction on an echocardiogram).

In comparing the relative value of the task definition and the algorithm, it could be argued that the value of the technical description of the underlying diagnostic task, which is based on extensive clinical experience and research, is greater than that of the algorithm. The algorithm is simply an automated instantiation of the technical description.

Classification of imaging findings of coronavirus disease 2019 presents an illustrative case study. Some authors simply relied on radiologists' judgment to evaluate presence or absence of disease [14-16]. A number of structured categorization schemes emerged to evaluate likelihood of disease, including (1) a consensus statement from the RSNA, which assigns an examination to one of the following categories: typical appearance, indeterminate appearance, atypical appearance, and negative for pneumonia [17], and (2) a classification scheme from a Dutch research group modeled on the ACR's BI-RADS to classify images on a scale from 0 to 6, with several categories matching or approximating those of the RSNA consensus statement [18]. Scoring systems for severity of disease were also developed, including (1) a 0 to 4 severity rating for each of six lung zones, for a total score of 0 to 24 [19], which was directly adapted from a system reported by Ooi et al in 2004 used to evaluate severe acute respiratory syndrome [20]; (2) a 0 to 5 severity rating for each of five lung lobes, for a total score of 0 to 25 by Pan et al [21], which was also directly adapted from a system used to evaluate severe acute respiratory syndrome [22], which itself was adapted from a system to evaluate interstitial lung disease [23]; and (3) a 0 to 7 severity rating for each of five lung lobes based on a combination of the approaches by Pan et al and Ooi et al, for a total score of 35 [20,21,24]. Any of these classification or scoring systems could be incorporated into an AI-based algorithm.

Conflating the task definition and the algorithm also conflates the responsibility to maintain and update the task definition. Although improving the performance of the algorithm is clearly the responsibility of the developer, it would be inappropriate for a developer to unilaterally maintain and update a widely utilized diagnostic task definition, such as the Bosniak classification scheme.

Superficial Treatment of the Diagnostic Task Definition

Numerous clinical assessment tasks, measurement systems, and classification schemes are amenable to automation by AI-based diagnostic algorithms (Tables 1 and 2). When these task definitions are used by manufacturers to design algorithms, they specify the diagnostic task that an algorithm performs. However, many of these task specifications were not developed using accepted consensus-based standard-setting processes [25] and are not maintained by dedicated standards bodies.

No Mechanism to Directly Compare Similar Algorithms

Because few of these defined diagnostic tasks were originally intended to be translated into AI-based diagnostic algorithms, they often lack detailed definitions and instructions for translation into algorithms. Thus, this important step is left to the interpretation of the algorithm developer, which ultimately limits the ability to directly compare the performance of AI-based diagnostic algorithms ostensibly developed for the same diagnostic task. This blunts the incentive for manufacturers to continuously strive for best-in-class performance.

Insufficient Characterization of Safety and Performance Elements

AI-based algorithms are prone to behaving in unpredictable ways when applied in the real world. For example, algorithm performance may degrade when applied to images generated by equipment from a different manufacturer or in a different clinical environment than those of the training set [26,27]. Algorithm performance can degrade over time when original training characteristics change [28]. Algorithms may return different outputs at different times when presented with almost exactly the same inputs [29,30]. Algorithm output may be affected by minor variations in image quality or extraneous data on

Table 1. Examples of types of measurements used as diagnostic tasks in medical imaging

Measurement Type	Example	Description
Lesion size	RECIST	Describes the technique for measuring lesion dimensions
Volumetric analysis	Future liver remnant	Quantifies volume of a lesion or organ of interest
Physiologic process	Gastric half emptying time	Quantifies the function of a physiologic process
Growth and maturity	Greulich and Pyle bone age	Quantifies growth and maturity of an individual or organ system
X-ray attenuation	Hounsfield unit	Quantifies attenuation of an x-ray beam on CT
Contrast enhancement	Peak lesion enhancement of on CT or MR	Quantifies enhancement based on threshold increase in attenuation or signal intensity after contrast administration CT or MR, respectively
Dynamic characteristics of contrast enhancement	Adrenal adenoma washout calculation	Describes the dynamics of decrease in lesion enhancements from the peak
Fat fraction	Liver fat fraction	Quantifies fat fraction within an organ or lesion based on MRI
Iron content	Liver T2* imaging	Quantifies iron content within an organ based on MRI
Diffusion-weighted imaging	Diffusion restriction	Quantifies Brownian motion of water molecules on MRI
Organ stiffness	Liver stiffness, based on MR or ultrasound elastography	Quantifies shear modulus of the liver on MR or ultrasound
Bolus perfusion	Stroke imaging	Quantifies mean transit time, cerebral blood volume, and cerebral blood flow to estimate tissue at risk
Fluid velocity	Portal vein velocity measure on ultrasound	Quantifies magnitude and direction of blood flow
Flow characteristics	Arterial resistive index on ultrasound	Estimates the resistivity to flow in an artery, based on minimum and maximum velocities
Radiotracer uptake	Standard uptake value	Ratio of radioactivity concentration in a lesion to whole body concentration of injected radiotracer

RECIST = Response Evaluation Criteria in Solid Tumours.

an image that is not salient to the diagnostic task [31-33]. Although regulatory review may prevent the most egregious currently known problems, it is unlikely to detect all problems or to drive vigorous research and innovation in evaluation methods.

Lack of Resources to Assess Performance at Each Installed Site

Algorithm performance tends to vary substantially from site to site in the real world [26,32,33]. This variability highlights the need for validation of algorithm performance at each clinical site before installation. However, full clinical validation is expensive and time-consuming. Current data science capabilities may not be scalable to meet users' demand for assurances of site-specific reliability and safety.

Inherent Conflicts of Interest

Manufacturers who develop and market SaMDs have a strong financial interest in showing their products in a positive light. Thus, an inherent conflict of interest exists if they are expected both to market their products and to fund, conduct, and publish results of objective and rigorous evaluation, including those results that may highlight deficiencies in their products.

RECOMMENDATIONS FOR IMPROVING THE CURRENT FRAMEWORK

The following recommendations address the described gaps.

Recommendation 1: Separate the Diagnostic Task From the Algorithm

Task definitions can be viewed as form of software specifications for SaMD applications based on human-labeled draining data. The algorithm's performance is inextricably linked to both the underlying task definition and the performance of the human labelers, reflecting the variability and biases in their understanding of and proficiency in the diagnostic task.

Because task definitions are a form of software specification, they should be developed according to accepted consensus-based standard-setting principles, which can help ensure widespread acceptance by relevant stakeholders [25]. As with any standard, diagnostic task definitions should be maintained by a nonconflicted entity committed to updating the definition based on new evidence and input from relevant stakeholders. Medical societies may be best suited for this task.

We propose that these task definitions should contain the following elements:

1. Background information, including a review of the evidence, the purpose of the task, all relevant definitions, and discussion of limitations and special cases;
2. A thorough description of the diagnostic task, including criteria for making the clinical assessment, descriptions and definitions of the measurement, or a description of all classification categories;
3. Detailed image labeling instructions for the task, including specific labeling strategies and relevant pitfalls;
4. Illustrated prototypical examples and relevant counter-examples, such as an atlas.

Medical societies may wish to go further and create a companion reference standard, which we define as a curated set of cases, carefully labeled based on the related diagnostic task definition, acquired from representative real-world populations in which the resulting algorithms will be used. Unlike an atlas, which is meant to be explanatory for human reviewers, the reference standard is designed to provide a common data set for objective assessment and comparison of algorithm performance in a controlled environment.

In the absence of a consensus definition of a diagnostic task, it is reasonable for an SaMD developer to develop an algorithm based on their best understanding of the diagnostic task. In fact, the developers may wish to propose and publish their own task definition. However, to become a true standard, the task definition must be adopted and maintained by a not-for-profit group with no financial interest in the development or marketing of any given algorithm.

To date, task definitions have typically been developed organically, with relatively little oversight and coordination. This has functioned reasonably well when algorithms are applied at the point of care by expert physicians who keep their knowledge current with the medical literature. As the value of the task definitions increases with the number of defined diagnostic tasks to which SaMD applications are applied, we foresee a need for active management of the ecosystem of task definitions.

In short, the advent of AI increases the urgency for medical professional societies to increase the development of rigorous, evidence-based diagnostic task definitions as consensus standards.

Recommendation 2: Define Performance Elements Beyond Accuracy

Perhaps the most concerning aspect of the unpredictable nature of AI-based diagnostic algorithms deployed in the clinical environment is the inability for algorithms to recognize and respond to these problems when they arise. Without internal monitoring mechanisms, algorithms will return erroneous information in the same way that they return correct information. This can lead to patient harm without the knowledge of the clinical team, local system administrators, or the manufacturer.

To address these risks, we recommend (1) the development of performance domains beyond accuracy, (2) the testing of all algorithms in all relevant domains before clinical deployment, and (3) the continuous monitoring of performance throughout the life cycle of the algorithm. A list of relevant performance domains is presented in Table 3.

Table 2. Examples of types of classification schemes used as diagnostic tasks in medical imaging

Classification Type	Example	Description
Findings associated with disease	Findings of pulmonary embolism on CT pulmonary angiogram	Describes the imaging findings associated with the presence (or absence) of a disease process
Probability of disease	PIOPED criteria	Provides a probability of disease based on imaging findings
Type of pathology	WHO CNS tumor classification system	Describes the type of disease, based on pathological provable characteristics, such as histopathology
Grade of pathology	Gleason grading system for prostate adenocarcinoma, AAST organ injury scoring scale	Describes the severity of disease, based on findings that are associated with relevant outcomes
Stage of pathology	TNM staging system	Describes the anatomical extent of the disease, based on findings that are associated with relevant outcomes
Lesion characterization based on risk assessment	BI-RADS	Categorizes lesions, or potential lesions, based on imaging findings that are associated with relevant outcomes
Pathology characterization	Stanford aortic dissection classification, Salter-Harris fracture types	Describes types of a disease process based on imaging findings and generally associated with relevant outcomes that may have implications for management
Diagnostic criteria based on imaging findings	Fleischner diagnostic HRCT criteria for UIP pattern	Image-based criteria for classifying a disease process
Clinical management based on imaging findings	Fleischner Society guidelines for management of incidental nodules detected on CT	Provides recommendations for clinical management based on imaging findings
Imaging pattern description	Wolfe classification of breast parenchymal patterns	Characterizes different types of normal imaging findings, typically to provide context for diagnosis of disease
Anatomical variants	Geist classification of os naviculare types	Describes different normal imaging appearances of an organ system

AAST = American Association for the Surgery of Trauma; CNS = central nervous system; HRCT = high resolution CT; PIOPED = Prospective Investigation of Pulmonary Embolism Diagnosis; UIP = usual interstitial pneumonia; WHO = World Health Organization.

It is especially critical that algorithms be tested for how they react when presented with unexpected data.

Recommendation 3: Divide the Evaluation Process Into Discrete Steps

We recommend dividing the characterization and evaluation of an AI-based diagnostic algorithm for medical imaging into the following steps. For each step, we pose questions from the perspective of a potential user or evaluator:

- Diagnostic task definition: What specific diagnostic task does the algorithm perform? Is the task defined in the same way for all algorithms purporting to perform the same task?
- Capability: How well does the algorithm perform its defined task in a controlled environment that simulates the real world, compared with other algorithms that perform the same task?
- Effectiveness (real-world performance): How does the algorithm perform, compared with its capability, when

Table 3. Examples of important performance elements of diagnostic algorithms

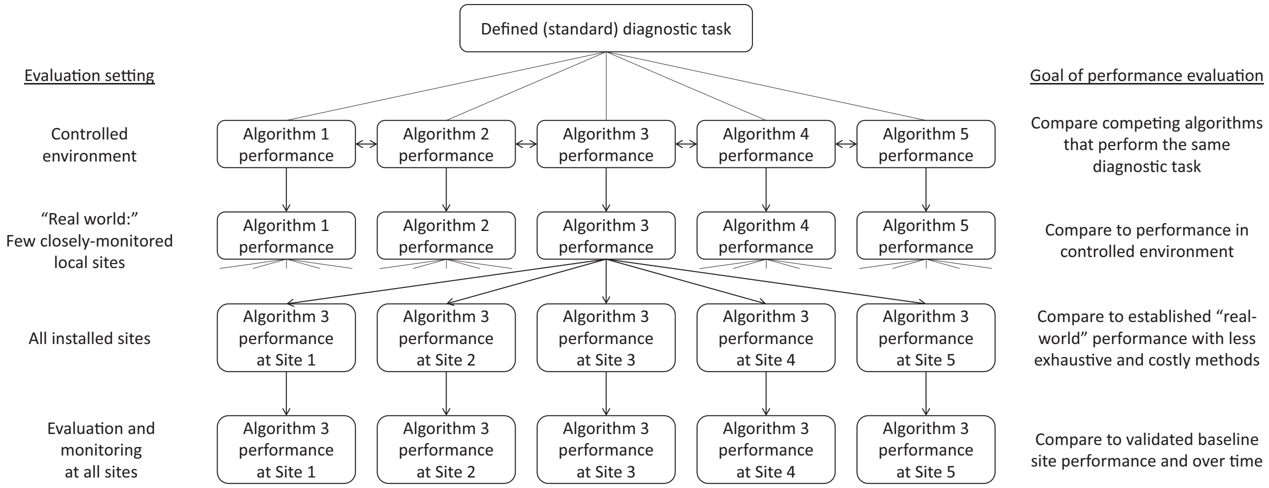
Element	Explanation
Accurate	The algorithm should accurately perform all diagnostic tasks for which it is designed.
Reliable	The algorithm should remain accurate in the setting of reasonably expected variation encountered in the clinical environment, including reasonable variations in image quality.
Applicable	The accuracy of the algorithm should be maintained across all makes and models of image modalities and for all patient populations for which it is designed to function.
Deterministic	The algorithm should give the same answer for the same image when used at different times and in different settings.
Nondistractible	The algorithm should be able to recognize the salient information from the image and not change its assessment based on extraneous, noncontributory image data.
Self-aware of limitations	The algorithm should have the means to detect when it is at or beyond the boundaries of its capabilities, whether because of inherent limitations of the model, limitations of its clinical applicability, or limitations imposed by clinical variation such as unexpected patient anatomy or image quality.
Fail-safe	The algorithm should recognize when it has reached an erroneous conclusion and have the means for ensuring that all errors are caught and stopped before they are propagated into the clinical environment.
Transparent logic	The user interface should enable the operator to clearly see the linkage between the input and output, including what data were analyzed, what alternatives were considered, and why certain possibilities were excluded, to be able to correctly accept or reject the algorithm's conclusion on any given case.
Transparent degree of confidence	The algorithm should share with the user a level of confidence in its assessment for each case. The accuracy of the model's expression of confidence should be validated as well as the accuracy of the model itself.
Able to be monitored	The algorithm should share performance data with users to enable ongoing monitoring of both individual and aggregated cases, quickly highlighting any significant deviations in performance.
Auditable	An independent means should be provided to monitor the algorithm's ongoing performance in a way that guides appropriate intervention. This may include periodic quality control checks similar to those performed by operators on imaging equipment.
Intuitive user interface	The user interface should enable the operator to intuitively how to use the algorithm with as little training as possible and impose the minimum possible cognitive load on the user.

deployed in a small number of closely monitored real-world settings?

- Effectiveness (local validity): How does the algorithm perform at every local site compared with its capability and established real-world performance at a few closely monitored sites?
- Durability: How does the algorithm perform over time, both in terms of maintaining and improving performance, with periodic updates as appropriate?

Each of these steps is dependent upon the successful completion of the previous step, as shown in Figure 1. For example, the diagnostic task must be well defined to provide a basis for comparative evaluation. Algorithms that do not perform well in a controlled environment are almost certain to not perform well in the real world. Algorithms that do not effectively perform in a few closely monitored real-world settings are unlikely to perform well at every installed site. If an algorithm's baseline performance at each local site cannot be determined, then its performance over time cannot be effectively monitored.

We believe that the comparison of algorithms performing the same task is best accomplished at the capability



Defined (standard) diagnostic task
Evaluation setting
Goal of performance evaluation
Controlled environment
Algorithm 1 performance
Algorithm 2 performance
Algorithm 3 performance
Algorithm 4 performance
Algorithm 5 performance
Compare competing algorithms that perform the same diagnostic task
"Real world:" Few closely-monitored local sites
Compare to performance in controlled environment
All installed sites
Algorithm 3 performance at Site 1
Algorithm 3 performance at Site 2
Algorithm 3 performance at Site 3
Algorithm 3 performance at Site 4
Algorithm 3 performance at Site 5
Compare to established "real-world" performance with less exhaustive and costly methods
Evaluation and monitoring at all sites
Compare to validated baseline site performance and over time

Fig 1. Illustration depicting proposed linkage of the evaluation of diagnostic algorithm performance from the defined task to implementation at the local site. Algorithms are developed according to a defined standard diagnostic task. Performance is compared with other algorithms in a controlled environment, which becomes the internal benchmark for general real-world performance and local site performance, which in turn becomes the benchmark for ongoing monitoring.

step, in which competing algorithms can be tested against a reference standard in a controlled environment along the performance dimensions listed in Table 3. This capability testing provides an external benchmark for the developer to strive for from the outset and an internal benchmark for the manufacturer to work to achieve across clinical implementation sites and over time. Through this process, the performance of deployed algorithms at all local sites is indirectly compared with that of the best-performing algorithms in the market (Fig. 1). To encourage a "race to the top," manufacturers should be required to share with users, prospective users, and regulators the results of performance testing at each step.

Because of the expense of conducting a full clinical evaluation at every site, once the algorithm's performance has been proven in a controlled environment and in the real world at a few closely monitored local sites, less intensive strategies for validation of local site performance should be sought that can provide reasonable assurance that local performance at a given clinical site is similar to real-world performance at other sites. Although such methods are not yet fully developed, one promising method is out-of-distribution detection, which assesses whether an image under question resembles images that the algorithm was trained on [34-37].

Recommendation 4: Encourage Assessment by Third-Party Evaluators

Despite their many positive elements, the described regulatory frameworks remain insufficient to incentivize and ensure excellence in every SaMD application at every site. For example, a manufacturer with a robust culture of quality and organizational excellence may use good machine learning practices to develop an algorithm that still has unrecognized shortcomings. Additionally, evaluation needs are likely to overwhelm regulatory resources. We believe that an objective third-party evaluator would be better suited to perform an exhaustive evaluation of the diagnostic algorithm according to the performance elements outlined in Table 3. Such third-party evaluators could include clinical research organizations, research laboratories, or organizations that develop and maintain reference standard data sets. Similar groups are commonly used to ensure quality data collection and best research practice in drug studies reviewed by the FDA [38,39].

Recommendation 5: Incorporate These Elements Into the Manufacturers' Development Process

SaMD applications are most likely to be successful if appropriate evaluation and improvement methods are incorporated into the software development process [6,11,12], which we describe in four phases: feasibility, capability, effectiveness, and durability (Fig. 2). These are analogous to development and evaluation phases that have been applied to pharmaceuticals [40] and proposed for software applications [41,42], as well as the design control phases highlighted in US and international regulatory documentation [43,44].

Phase I: Feasibility. In this first phase, developers train an algorithm to classify a data set according to the defined diagnostic task. The goal is to demonstrate whether the

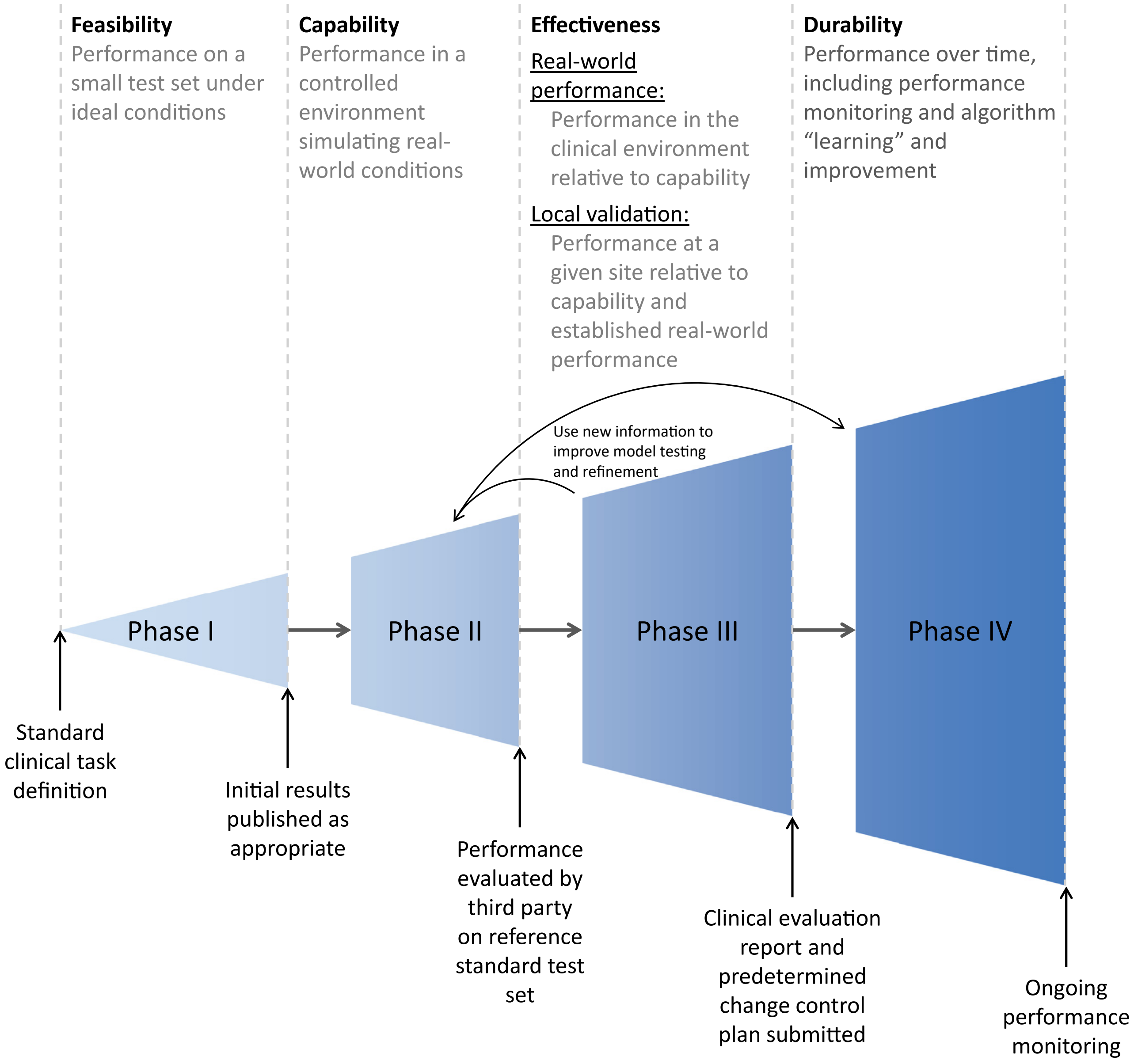

Feasibility
Performance on a small test set under ideal conditions
Capability
Performance in a controlled environment simulating real-world conditions
Effectiveness
Real-world performance:
Performance in the clinical environment relative to capability
Local validation:
Performance at a given site relative to capability and established real-world performance
Durability
Performance over time, including performance monitoring and algorithm "learning" and improvement
Use new information to improve model testing and refinement
Phase I
Phase II
Phase III
Phase IV
Standard clinical task definition
Initial results published as appropriate
Performance evaluated by third party on reference standard test set
Clinical evaluation report and predetermined change control plan submitted
Ongoing performance monitoring

Fig 2. Phased development and evaluation process of diagnostic algorithms.

algorithm can compete with the current state of the art (which may be existing algorithms or expert humans) under ideal conditions on at least a single cohort of clinical images. In this manner, the specification requirements (design input) [45] of a system are documented at a product level, and feasibility assessments are performed against them.

Algorithms do not need to be fully robust at this stage, because the goal is simply to demonstrate feasibility. The resulting findings may be worthy of publication [46], although the algorithm is not yet ready for full deployment in the clinical environment.

Phase II: Capability. In this phase, developers refine the algorithm on broadly representative input data until they can demonstrate that the algorithm consistently performs as intended in an environment simulating real-world conditions. The algorithm's accuracy, reliability, and safety should be objectively measured against all of the criteria listed in Table 3. In this context, "accuracy" refers to measures of how closely the algorithm's output (or the design output of a system) [45] matches the ground truth, including sensitivity, specificity, and predictive value. "Reliability" refers to the algorithm's ability to consistently perform accurately in all conditions under which it may be used. "Safety" refers to the algorithm's ability to minimize the risk of harm when deployed, including when subjected to unanticipated situations.

Algorithms should be developed following principles promoting quality and safety for SaMDs, including risk management (or safety), quality management, and systems engineering according to best practices [5,43,47,48]. The algorithm should be evaluated according to how it interfaces with the operator, including the cognitive burden placed upon the individual, and how well the algorithm and the user perform together in the typical environment. This evaluation has historically been performed via reader and user studies [5,49].

The algorithm should be deliberately stress tested for design verification [45] in suboptimal conditions, simulating real-world conditions when possible.

Although this phase involves testing on retrospective data, the controlled environment is the defining feature at this stage rather than the retrospective or prospective nature of the data. The algorithm developer should continue to maintain a high-fidelity test environment during the remaining phases and throughout the deployment life cycle of the algorithm. As later phases of algorithm testing and deployment yield new information, both the algorithm and the controlled testing environment should be updated accordingly.

Before proceeding to deployment in the clinical setting, the algorithm should be evaluated by a third party on a reference standard test set, incorporating testing on all relevant performance domains, such as those listed in Table 3.

Phase III: Effectiveness. Assessment of effectiveness can be considered in two parts: general real-world performance and local validation.

General real-world performance or design verification can be determined before full clinical deployment by prospectively evaluating the algorithm in at least a few closely monitored real-world clinical settings. The primary objective is to confirm that the real-world performance of the algorithm matches its performance in the test environment [7, 50]. All learnings from this stage should be incorporated into the algorithm, which must be retested in the controlled environment before being updated in the real world.

Local validation should be performed by the manufacturer at each site before or at the time of clinical implementation. Manufacturers may utilize validation methods that are less exhaustive and costly than full evaluation of general real-world performance, but those methods should be validated before fully marketing the SaMD.

Real-world deployment may reveal quality control problems at the local clinical sites [51]. Manufacturers should prepare strategies to distinguish whether performance problems are due to suboptimal algorithm performance versus local quality control problems and work with local clinical sites to help resolve quality issues [51].

At this stage, the manufacturer should submit required documentation, such as the clinical evaluation report in the European Union [52] and the predetermined change control plan, according to the proposed FDA framework [12].

Phase IV: Durability. Under the total product life cycle concept, manufacturers have an obligation to support the product throughout its clinical implementation. This includes ongoing performance evaluation and monitoring, with the intent of continuous improvement [11,12]. The IMDRF recommends that manufacturers embed monitoring or auditing systems within their products to automatically detect, recover from, and report errors [5]. They should also seek less structured sources of feedback, including customer inquiries, complaints, market studies, focus groups, and field service reports [5].

When problems are encountered following algorithm deployment, algorithms should be modified and thoroughly tested in the controlled environment before they are reintroduced as a new SaMD version. Algorithms should also be regularly updated as improvements become available, such as advances in technology, updates in task definitions, or better training data. Each new version must be carefully tracked and closely monitored with strict change control procedures, including a reference to the version of classification systems or other diagnostic task definitions [5,10].

We strongly agree with the IMDRF that manufacturers should actively monitor the technical performance of the algorithm on its defined diagnostic task [5,7]. However, we believe that evaluation of the clinical effectiveness of the task definition need not be performed by the manufacturer, as long as the task definition adheres to published definitions based on clinical research or professional society guidelines.

In conclusion, although device regulators have provided a strong foundation for ensuring the quality and safety of SaMD applications, we have outlined recommendations to fill gaps in these frameworks that may otherwise prevent the development of a healthy AI ecosystem that creates a race to the top in terms of accuracy, reliability, and safety.

The need for rigorous, evidence-based diagnostic task definitions exists independent of SaMD applications, but its urgency is dramatically increased by the ability to scale automated diagnosis through AI-based diagnostic algorithms.

We anticipate that the growing number and type of algorithms will drive the emergence of a substantial body of research dedicated to ensuring the accuracy, reliability, and safety of the algorithms—possibly rivalling the body of research dedicated to the development of the algorithms themselves.

Ensuring that diagnostic algorithms perform effectively both in controlled environments and in the real world can help ensure the medical community and the public the algorithms have been thoroughly tested and refined before being deployed in the clinical environment, just as they have come to expect from other medical devices.

TAKE-HOME POINTS

- Strategies outlined by regulatory bodies address many key aspects to help ensure the safety, effectiveness, and performance of SaMD applications, but a number of gaps remain.
- Appropriate evaluation and improvement methods should be incorporated into phases of development, analogous to those that have been applied to pharmaceuticals and proposed for software applications.
- Algorithms should be thoroughly tested and refined before being deployed in the clinical environment, just as has come to be expected from other medical devices.
- Regulatory frameworks should strive to establish conditions that set up a race to the top for consistent excellent algorithm performance at each installed site.

REFERENCES

1. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018;2:35.
2. Roski J, Chapman W, Heffner J, et al. How artificial intelligence is changing health and health care." in artificial intelligence in health care: the hope, the hype, the promise, the peril. Washington, DC: National Academy of Medicine; 2019.
3. The Lancet. Is digital medicine different? Lancet 2018;392:95.
4. IMDRF Software as a Medical Device (SaMD) Working Group. Software as a medical device (SaMD): key definitions. Doc no. IMDRF/SaMD WG/N10 FINAL:2013. International Medical Device Regulators Forum (IMDRF). Available at: http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf. Published September 9, 2013. Accessed January 25, 2019.
5. IMDRF Software as a Medical Device (SaMD) Working Group. "Software as a medical device": possible framework for risk categorization and corresponding considerations. Doc no. IMDRF/SaMD WG/N12 FINAL:2014. International Medical Device Regulators Forum (IMDRF). http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf. Accessed January 4, 2021.
6. IMDRF Software as a Medical Device (SaMD) Working Group. Software as a medical device (SaMD): application of quality management system. Doc no. IMDRF/SaMD WG/N23 FINAL: 2015. International Medical Device Regulators Forum (IMDRF). http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-151002-samd-qms.pdf. Accessed January 4, 2021.
7. IMDRF Software as a Medical Device (SaMD) Working Group. Software as a medical device (SaMD): clinical evaluation. Doc no. IMDRF/SaMD WG/N41 FINAL:2017. International Medical Device Regulators Forum (IMDRF). http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf. Accessed January 4, 2021.
8. Medical Device Clinical Evaluation Working Group. Clinical evaluation. Doc no. IMDRF/MDCE WG/N56 FINAL:2019. International Medical Device Regulators Forum (IMDRF). http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf. Accessed January 4, 2021.
9. Medical Device Clinical Evaluation Working Group. Clinical investigation. Doc no. IMDRF/MDCE WG/N57 FINAL:2019. International Medical Device Regulators Forum (IMDRF). http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n57.pdf. Accessed January 3, 2021.
10. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745. Approved April 5, 2017. Accessed January 25, 2019.
11. US FDA. Developing a software precertification program: a working model. Version 1.0. Available at: https://www.fda.gov/media/119722/download. Published January 2019. Accessed January 25, 2019.
12. US FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. Available at: https://www.fda.gov/media/122535/download. Published April 2, 2019. Accessed January 25, 2019.
13. Bosniak MA. The current radiological approach to renal cysts. Radiology 1986;158:1-10.
14. Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020;296:E115-7.
15. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020;296(2):E32-40.
16. Caruso D, Zerunian M, Polici M, et al. Chest CT features of COVID-19 in Rome, Italy. Radiology 2020;296:E79-85.
17. Simpson S, Kay FU, Abbara S, et al. Radiological Society of North America expert consensus statement on reporting chest CT findings related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA -Secondary Publication. J Thorac Imaging 2020;35:219-27.
18. Prokop M, van Everdingen W, van Rees Vellinga T, et al. COVID-19 Standardized Reporting Working Group of the Dutch Radiological Society. CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19: definition and evaluation. Radiology 2020;296:E97-104.
19. Wang Y, Dong C, Hu Y, et al. Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: a longitudinal study. Radiology 2020;296:E55-64.
20. Ooi GC, Khong PL, Müller NL, et al. Severe acute respiratory syndrome: temporal lung changes at thin-section CT in 30 patients. Radiology 2004;230:836-44.
21. Pan F, Ye T, Sun P, et al. Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia [E-pub ahead of print, 2020 Feb 13]. Radiology 2020:200370.
22. Chang YC, Yu CJ, Chang SC, et al. Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT. Radiology 2005;236:1067-75.
23. Kazerooni EA, Martinez FJ, Flint A, et al. Thin-section CT obtained at 10-mm increments versus limited three-level thin-section CT for idiopathic pulmonary fibrosis: correlation with pathologic scoring. AJR Am J Roentgenol 1997;169:977-83.
24. Huang G, Gong T, Wang G, et al. Timely diagnosis and treatment shortens the time to resolution of coronavirus disease (COVID-19)

pneumonia and lowers the highest and last CT scores from sequential chest CT. AJR Am J Roentgenol 2020;215:367-73.

25. American National Standards Institute. About ANSI. Introduction. Available at: https://www.ansi.org/about_ansi/introduction/introduction. Accessed May 10, 2020.

26. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. J Am Coll Radiol 2020;17:796-803.

27. Subbaswamy A, Schulam P, Saria S. Preventing failures due to dataset shift: Learning predictive models that transport. Proc Mach Learn Res 2019;89:3118-27.

28. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020;21:345-52.

29. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195.

30. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 2019;155:1135-41.

31. Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. Science 2019;363:1287-9.

32. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683.

33. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci U S A. 2020 May; pii: 201907377. https://doi.org/10.1073/pnas.1907377117. Epub ahead of print. PMID: 32393633.

34. Lee K, Lee H, Lee K, Shin J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv:1711.09325 Available at: https://arxiv.org/abs/1711.09325 Submitted November 26, 2017. Updated February 23, 2018. Accessed May 10, 2020.

35. DeVries T, Taylor GW. Learning confidence for out-of-distribution detection in neural networks. arXiv:1802.04865 Available at: https://arxiv.org/abs/1802.04865 Submitted Feb 13, 2018. Accessed May 10, 2020.

36. Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv:1706.02690 Available at: https://arxiv.org/abs/1706.02690 Submitted June 8, 2017. Updated February 25, 2018. Accessed May 10, 2020.

37. Ren J, Liu PJ, Fertig E, Snoek J, Poplin R, DePristo MA, Dillon JV, Lakshminarayanan B. Likelihood ratios for out-of-distribution detection. arXiv:1906.02845 Available at: https://arxiv.org/abs/1906.02845 Submitted June 7, 2019. Updated December 5, 2019. Accessed May 10, 2020.

38. Hecker SJ, Preston C, Foote M. Production of high-quality marketing applications: strategies for biotechnology companies working with contract research organizations. Biotechnol Annu Rev 2003;9:269-77.

39. Beckett M, Quiter E, Ryan G, et al. Bridging the gap between basic science and clinical practice: the role of organizations in addressing clinician barriers. Implement Sci 2011;6:35.

40. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: a narrative review. Postgrad Med 2011;123:194-204.

41. Spiegelhalter DJ. Evaluation of clinical decision-aids, with an application to a system for dyspepsia. Stat Med 1983;2:207-16.

42. Stead WW, Haynes RB, Fuller S, et al. Designing medical informatics research and library—resource projects to increase what is learned. J Am Med Inform Assoc 1994;1:28-33.

43. International Organization for Standardization (ISO) Technical Committee ISO/TC 210 Quality Management and Corresponding General Aspects for Medical Devices. IEC 62304:2006 Medical device software—software life cycle processes. ISO. Available at: https://www.iso.org/standard/38421.html. Published May 2006. Accessed January 25, 2019.

44. US FDA. Quality system regulation. 21 CFR §820. Available at: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=820. Accessed January 25, 2019.

45. US FDA. Quality system regulation. 21 CFR §820.30(c). Available at: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=820.30. Accessed January 25, 2019.

46. Bluemke DA. Radiology in 2018: are you working with AI or being replaced by AI? Radiology 2018;287:365-6.

47. International Organization for Standardization (ISO) Technical Committee ISO/TC 210 Quality Management and Corresponding General Aspects for Medical Devices. ISO 13485:2016 Medical devices—quality management systems—requirements for regulatory purposes. ISO Available at: https://www.iso.org/standard/59752.html Published September 2015. Accessed January 25, 2019.

48. International Organization for Standardization (ISO) Technical Committee ISO/TC 210 Quality Management and Corresponding General Aspects for Medical Devices. ISO 14971:2019 Medical devices—application of risk management to medical devices. ISO. Available at: https://www.iso.org/standard/72704.html. Published December 2019. Accessed May 10, 2020.

49. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) Joint Technical Committee ISO/IEC JTC1, Information Technology, Subcommittee SC 27, IT Security Techniques. ISO/IEC 27001:2013—Information technology—security techniques—information security management systems—requirements. ISO. Available at: https://www.iso.org/standard/54534.html. Published October 2013. Accessed January 25, 2019.

50. International Organization for Standardization (ISO) Technical Committee ISO/TC 194 Biological and clinical evaluation of medical devices. ISO 14155:2011 Clinical investigation of medical devices for human subjects—good clinical practice. ISO. Available at: https://www.iso.org/standard/45557.html. Published February 2011. Accessed May 10, 2020.

51. Larson DB, Boland GW. Imaging quality control in the era of artificial intelligence. J Am Coll Radiol 2019;16(9 Pt B):1259-66.

52. European Commission (DG Internal Market, Industry, Entrepreneurship and SMEs Consumer, Environmental and Health Technologies Health technology and Cosmetics). Guidelines on medical devices. MEDDEV 2.7/1 revision 4. Available at: https://ec.europa.eu/docsroom/documents/17522/attachments/1/translations/en/renditions/native. Published June 2016. Accessed May 10, 2020.