JMI Journal of Medical Imaging

REVIEW

Regulatory considerations for medical imaging AI/ML devices in the United States: concepts and challenges

Nicholas Petrick,[a,*] Weijie Chen,[a] Jana G. Delfino,[a] Brandon D. Gallas,[a] Yanna Kang,[b] Daniel Krainak,[b] Berkman Sahiner,[a] and Ravi K. Samala[a]

[a]U.S. Food and Drug Administration, Center for Devices and Radiological Health, Office of Science and Engineering Labs, Silver Spring, Maryland, United States

[b]U.S. Food and Drug Administration, Center for Devices and Radiological Health, Office of Product Evaluation and Quality, Silver Spring, Maryland, United States

ABSTRACT. **Purpose:** To introduce developers to medical device regulatory processes and data considerations in artificial intelligence and machine learning (AI/ML) device submissions and to discuss ongoing AI/ML-related regulatory challenges and activities.

Approach: AI/ML technologies are being used in an increasing number of medical imaging devices, and the fast evolution of these technologies presents novel regulatory challenges. We provide AI/ML developers with an introduction to U.S. Food and Drug Administration (FDA) regulatory concepts, processes, and fundamental assessments for a wide range of medical imaging AI/ML device types.

Results: The device type for an AI/ML device and appropriate premarket regulatory pathway is based on the level of risk associated with the device and informed by both its technological characteristics and intended use. AI/ML device submissions contain a wide array of information and testing to facilitate the review process with the model description, data, nonclinical testing, and multi-reader multi-case testing being critical aspects of the AI/ML device review process for many AI/ML device submissions. The agency is also involved in AI/ML-related activities that support guidance document development, good machine learning practice development, AI/ML transparency, AI/ML regulatory research, and real-world performance assessment.

Conclusion: FDA's AI/ML regulatory and scientific efforts support the joint goals of ensuring patients have access to safe and effective AI/ML devices over the entire device lifecycle and stimulating medical AI/ML innovation.

Keywords: AI/ML; regulatory concepts; medical imaging; assessment methods; regulatory science

Paper 23031SSR received Feb. 2, 2023; revised May 22, 2023; accepted May 30, 2023; published Jun. 23, 2023.

1 Introduction

Devices incorporating artificial intelligence and machine learning (AI/ML) can be found in many areas of medicine, and the U.S. Food and Drug Administration (FDA) and Center for Devices and Radiological Health (CDRH) have a history regulating medical AI/ML technologies.[1] As an example, a semi-automated cervical cytology slide reader incorporating neural network

*Address all correspondence to Nicholas Petrick, Nicholas.Petrick@fda.hhs.gov

processors was first approved by the FDA in 1995.[2] FDA receives a high volume of premarket submission inquiries for products leveraging AI/ML technologies, and we expect this trend to continue going forward. Q-submissions are a mechanism for device developers to ask questions and obtain official FDA feedback prior to a formal premarket submission.[3] Q-submissions are highly recommended to help address a question prior to a full device submission.

The International Medical Device Regulators Forum defined software as a medical device (SaMD)[4] as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device." While not all SaMD incorporates AI/ML and not all devices involving AI/ML are SaMD, many AL/ML devices are developed and implemented independently from the image acquisition and display devices such that they fit under the wider SaMD umbrella.

Radiology has been a pioneer in adopting AI/ML-enabled devices into the clinical environment. Radiological AI/ML devices are numerous and expanding with applications aiming to improve the efficiency, accuracy, or consistency of the medical image interpretation process across a wide range of radiological tasks and imaging modalities. A non-exhaustive list of AI/ML-enabled medical devices authorized for marketing in the United States and identified through FDA's publicly available information can be found on the FDA webpage.[5] Table 1 defines some common types or classes of medical imaging AI/ML that have been submitted to the FDA. Each product type, and its associated product code, contains one or more devices authorized for marketing in the United States with most of these devices considered SaMD.

AI/ML-enabled medical devices present many opportunities for improving medical practice through their ability to learn from real-world data, improve performance over time, and simplifying the image interpretation process for the clinician by automating routine and tedious tasks. However, these devices come with unique challenges, including the need for large and representative datasets, propagation of biases within the training data and data sourcing process, understanding the upstream and downstream role and impact on clinical workflows, and

Table 1 A cross section of device product codes that included medical imaging AI/ML. Each product type, and its associated product code, contains at least one AI/ML medical device authorized for marketing in the United States.

FDA product code	Device	Short definition
OMJ[6]	Chest x-ray computer aided detection	Software device to assist radiologists in the review of chest radiographic images and highlight potential nodules that the radiologist should review
QDQ[7]	Radiological computer assisted detection/diagnosis software for lesions suspicious for cancer	An image processing device intended to aid in the detection, localization, and characterization of lesions suspicious for cancer on acquired medical images (e.g., mammography, MR, CT, ultrasound, radiography)
QAS[8]	Radiological computer-assisted triage and notification software	An image processing device intended to aid in prioritization and triage of time-sensitive patient detection and diagnosis based on the analysis of medical images acquired from radiological signal acquisition systems
QJU[9]	Image acquisition and/or optimization guided by artificial intelligence	A device that is intended to aid in the acquisition and/or optimization of images and/or diagnostic signals
QNP[10]	Gastrointestinal lesion software detection system	A computer-assisted detection device used in conjunction with endoscopy for the detection of abnormal lesions in the gastrointestinal tract
QPN[11]	Software algorithm device to assist users in digital pathology	An *in vitro* diagnostic device intended to evaluate acquired scanned pathology whole slide images

difficulties assuring continued safety and effectiveness over time for fixed models, e.g., because of changes to the clinical population, and for continuously learning AI/ML devices.

The agency is striving to address challenges around AI/ML development and assessment to ensure patients have access to safe and effective AI/ML devices. In addition to a high-level white paper on good machine learning practices,[12] the agency is developing regulatory policies,[13] conducting regulatory science research, and collaborating with stakeholders to better understand and characterize AI/ML models and develop least-burdensome assessment methods.[14] Similarly, other groups and organizations are working to develop consensus best practices for medical imaging AI/ML.[15–18] Some of the more wide-ranging efforts specific to medical imaging AI/ML include the FUTURE-AI guiding principles developed by five European AI in Health Imaging projects[15] and the American Association of Physicists in Medicine Task Group Report 273 discussing best practices for medical imaging computer-aided diagnosis.[16]

In this review paper, we introduce the reader to the medical device regulatory framework within the United States and provide an overview of common elements included in regulatory submissions that incorporate AI/ML models in medical imaging in Sec. 2.1. Specifically, we discuss the model description, data, nonclinical testing and multi-reader, multi-case studies used to evaluate the device in the hands of the end user in Secs. 2.2–2.5. Finally, we discuss ongoing and planned activities adapting FDA regulatory processes to AI/ML device submissions and how the agency is addressing regulatory science gaps in Sec. 3 of this paper.

2 Methods

2.1 Regulatory Framework

The FDA regulates medical device manufacturers based on the level of risk posed by the device, which is informed by the intended use of the device, the indications for use of the device, and the technological characteristics of the device. The intended use describes the general purpose of the device or its function while the indications for use are more specific, describing the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended.[19] The intended use is important in determining both the regulatory pathway and what data and information are necessary in a regulatory submission of that device.

2.1.1 *Product classification and regulatory controls*

The FDA classifies medical devices into classes I, II, or III, where the class is based on the device risk and determines the extent of regulatory controls necessary to provide a reasonable assurance of the safety and effectiveness for the device (21 C.F.R. § 860). All devices, regardless of class, must adhere to the general control provisions of the Food, Drug, and Cosmetic Act that relate to adulteration; misbranding; device registration as well as device listing; premarket notification, banned devices; notification, including repair, replacement, or refund; records and reports; restricted devices; and good manufacturing practices.[20] Most medical image processing devices, with or without AI/ML, are currently classified as class II. Aside from some exemptions and in addition to general controls, to market a class II device, a manufacturer must describe and test their device according to all applicable special controls[21] and demonstrate substantial equivalence between their new device and a legally marketed device, i.e., the predicate device.[19] Substantial equivalence is established with respect to intended use, design, energy used or delivered, materials, performance, safety, effectiveness, labeling, biocompatibility, standards, and other applicable characteristics in a Class II Premarket Notification [510(k)] submission. Many of these characteristics are described in the product classification of the predicate.[22] If the device is determined to be substantially equivalent, the manufacturer is then "cleared" to market that device in the United States.

2.1.2 *De Novo pathway*

The De Novo pathway provides a pathway to a class I or class II classification for medical devices for which general controls or general and special controls provide a reasonable assurance of

safety and effectiveness, but for which there is no legally marketed predicate device.[23] This pathway was created as an alternative to the default high-risk class III classification. Granting of a De Novo establishes a new regulation not only for the specific device but also more broadly for the general device type and intended use, which will then be used for regulation of subsequent devices, e.g., only the first one is a De Novo. Multiple AI/ML device types have been classified as class II through the De Novo pathway, establishing specific, published special controls for the appropriate product class. Example product classes are computer-aided detection (CADe) for lesions using optical colonoscopy[24] and radiological computer-aided diagnosis (CADx) for lesions suspicious for cancer.[25] De Novo submissions can also establish broader product classes, as with the first digital pathology AI/ML device, which is more broadly defined as "software algorithms to provide information to the user about presence, location, and characteristics of areas of the image with clinical implications. Information from this device is intended to assist the user in determining a pathology diagnosis."[26]

2.1.3 *Predicates*

Choosing a predicate device for a 510(k) submission is an important first step in pursuing marketing clearance for a class II device. As mentioned, the predicate establishes the product classification, special controls, and the safety and effectiveness information necessary to determine substantial equivalence. As such, we recommend manufacturers engage with CDRH reviewers early through the Q-submission program[3] when identifying a predicate is not clear or when they have other review process questions. The Q-submission process can allow manufacturers to determine if the FDA agrees with their suggested product class or predicate and can be used to address specific questions a manufacturer may have about the testing necessary to demonstrate safety and effectiveness.

2.1.4 *AI/ML premarket submissions*

Depending on the product class and technologies incorporated into the overall device, premarket AI/ML device submissions can contain a wide array of information and testing to facilitate the review process, e.g., a description of the device, a discussion of relevant standards conformed with, nonclinical studies documentation and testing, clinical studies documentation and testing, software documentation and testing, and cyber-security documentation and testing. Figure 1 shows some of the common considerations in AI/ML device assessment to be addressed as part of an AI/ML device submission. We will not discuss all the aspects of a premarket AI/ML device

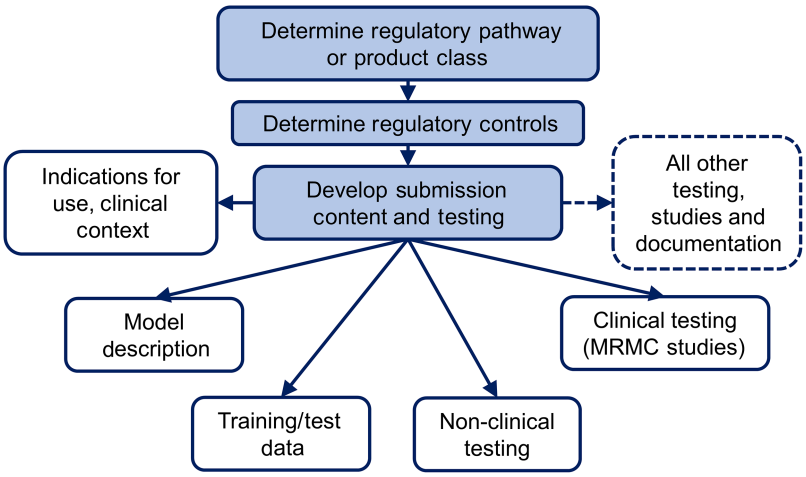

Fig. 1 Flowchart depicting some common considerations for AI/ML device evaluation as part of a premarket submission.

submission here, but we will discuss four critical aspects including model description, data, nonclinical testing, and multi-read multi-case testing.

2.2 Model Description

A conceptual description of an AI/ML model is a great start, but an engineering description of the architecture and how it was built is better. The description can include references to literature and figures with flow charts and diagrams. A detailed engineering description of the software helps reviewers understand the underlying functionality and complexity of the device, determine how it should be tested, and whether the performance of the device is expected to generalize to different data acquisition devices or patients. Details on the following are generally important to describe:

- input data and the dimension of each input, including patient images and patient meta data,
- engineered features and the feature selection processes, if appropriate,
- pre-processing and post-processing necessary for AI/ML model application,
- model network type(s) and components, e.g., model architecture including layers, activation functions, loss function, and the dimensionality of the data throughout the processing pipeline, and
- model development, including training and tuning processes, e.g., transfer learning, data augmentation, regularization methods, ensemble methods, tuning thresholds and hyperparameters, optimization methods (optimality criteria), performance assessment metrics, calibration, and other documentable parameters.

2.3 Data

The data used in AI/ML training and testing are critical for developing robust models, especially when implementing a deep learning method that combines and automates feature extraction, feature selection, and classification. We refer to the patient images together with other patient and clinical information as data.[27,28] Data for AI/ML device submissions come from a number of different sources including data collected by the manufacturer specifically for AI/ML device development and evaluation, other private data collection efforts, public data collections, and potentially even via synthetic data. Each of these collection sources has unique challenges and benefits related to burden, quality of the reference standard, representing the intended patient population, representing the intended image acquisition devices, and controlling access, e.g., preventing commingling of training and test data. The choice of data for any specific use or application is a tradeoff between these various factors. When collecting image data for developing or testing AI/ML models, it is important to also acquire appropriate clinical information, e.g., patient demographics, family history, reason for the exam; disease specific information, e.g., disease type and lesion size; image acquisition information, e.g., patient prep, device manufacturer and model, protocol, and reconstruction method; and other clinical test results in order to characterize and understand training and testing limitations and model generalizability. Providing tables and diagrams characterizing the data helps facilitate an efficient review process. Since developers often explore multiple model architectures and fine-tune parameters as part of the training process,[29,30] it is useful to provide a flow chart of the entire development process and clarify the methods and data used in each step.

Dataset size is an important resource allocation issue. A general principle is for the training dataset to be large enough to allow an AI/ML model to learn the relationship between the input and output with little or no overfitting, and for the independent test dataset to be large enough to provide adequate precision of the performance estimates, i.e., sufficiently small error bars. The datasets, both training and test, should include relevant cohorts and subgroups containing enough patients/cases to facilitate robust algorithm training and facilitate subgroup analyses as discussed in Sec. 2.4. Research has shown that as the training set is gradually increased from a small size, overfitting initially decreases dramatically, with diminishing returns as the dataset size gets larger.[31] The rate at which adding more data improves performance depends on the complexity of the AI/ML model and the complexity of the data space. Estimation of the test dataset size for adequate precision and study power is a classical problem in statistics, and pilot data are extremely helpful for estimating the dataset sizes appropriately.

2.3.1 *Independence*

A central principle for performance evaluation is that the test dataset should be independent of the training dataset (different patients and different clinical sites) to avoid biases in performance assessment and demonstrate performance generalizability. Violation of this principle will result in optimistically biased performance estimates and potentially unacceptable real-world performance.[32] Although this principle is simple, there are subtle ways in which the independence principle can be violated. A typical example of violation arises from the failure to recognize that multiple images (or image regions) from the same patient are correlated. The images from the same patient are different, but not independent. Therefore, to avoid information leakage between the training and testing datasets, the images from each patient should appear in only one of these datasets.

A more subtle mechanism that can cause dependence between the training and test datasets arises when a single-site dataset is randomly split into training and testing datasets. This approach seems logical, but it may result in a site-specific similarity between the training and test datasets. Essentially, the training and testing data from the single site are more alike than what might be found in the real world such that this approach can underestimate overall variability, e.g., cross-site variability in data acquisition and clinical practice. When this approach is used, it is referred to as internal validation.[33] The limitation of this approach can be mitigated by including data from multiple sites, but it is better to split the training and test data by site or at least by time. This approach is referred to as external validation.[34] Figure 2 shows the need for the AI/ML development to be independent from the performance assessment conducted as part of a device submission including having the test data site and time independent from that of the AI/ML training and tuning data.

2.3.2 *Representativeness*

The collected data should be large in size and representative of the target population. One approach is to collect consecutive cases that fit the intended use population from a diverse set of sites, e.g., academic, community, rural health providers, over a defined time period. This approach can be resource intensive when disease prevalence is low. Therefore, an alternative approach may be necessary, but it should still address the representativeness of the collected data as much as possible.

When collecting a training dataset, a key is to include a diverse set of cases. It may be efficient (improve model performance) to consider sampling methods that target "informativeness," "representativeness," or a combination of the two.[35,36] For a testing dataset, it is more

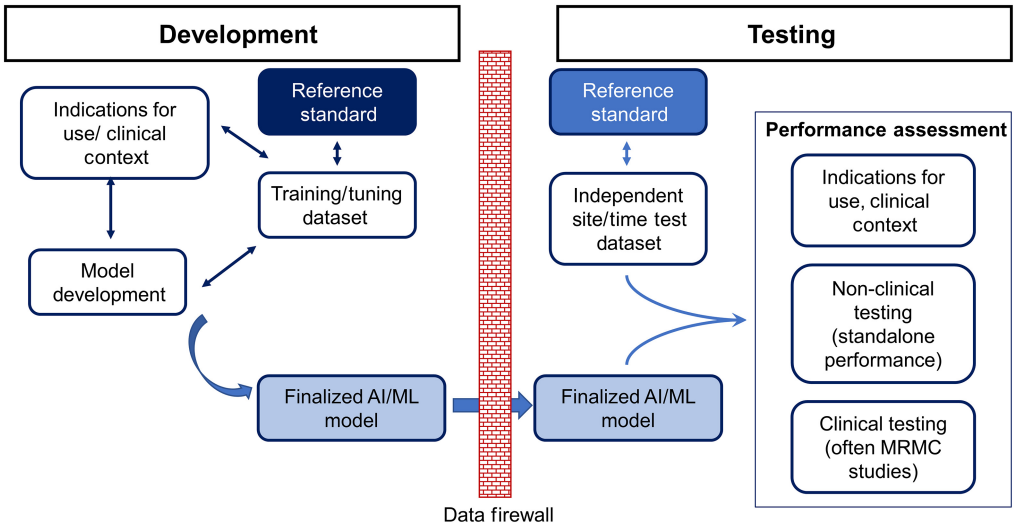

Fig. 2 Flowchart depicting the AI/ML development process and its independence from the performance assessment conducted as part of a device submission. The test data are ideally site and time independent from that of the AI/ML training and tuning data.

important to match the study data with the target population, but there may be some flexibility under FDA's least burdensome principle[37] in studies with controlled design and informed interpretation of results. It may be necessary to include rare cases or patient subgroups, or it may be statistically efficient to stratify, or enrich, the sampling across patient subgroups. If the enrichment is based solely on the disease condition, the impact may only be on the prevalence of the study set. Several performance metrics are unaffected by prevalence differences, e.g., sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve, denoted AUROC. Other differences between the study and clinical population can result in differences between reported device performance in the study and true performance on the clinical population, which may be problematic if not addressed. Questions about the appropriateness of testing datasets and assessment protocols are best addressed on a case-by-case basis in a Q-submission.

2.4 Nonclinical Testing, Standalone Performance

Nonclinical testing is a catchall category for performance testing that is not related to the active observation and treatment of patients.[38] This includes mechanical, biological, and engineering performance (fatigue, wear, tensile strength) using *ex vivo*, *in vitro*, *in situ*, and simulation studies. For AI/ML devices, nonclinical testing may include software verification and validation, human factors validation, and standalone performance among other tests depending on the device and application area. Standalone performance is a measure of device performance alone with little to no interaction or interpretation from a clinical end user.[28] When a clinical end user needs to interact with the device or interpret the device outputs, FDA generally requires an assessment of the device in the hands of the end users. Such an assessment happens in a "reader study," which is discussed in Sec. 2.5.

Standalone testing also provides a performance benchmark for comparing AI/ML devices from the same or different manufacturers. This benchmark can reduce the need for clinical performance testing in future regulatory submissions. Putting aside the challenges related to establishing the reference standard, standalone testing is largely a computational exercise and can be systematically applied to a large number of sample cases. As such, standalone testing can be very useful for assessing device generalizability to different clinical subpopulations, sites, and to different image acquisition devices and protocols.[39,40]

2.4.1 *Evaluation metrics*

Selection of performance metrics is crucial for benchmarking an AI/ML model and for comparing performance with predicate devices or other appropriate comparators. The endpoints selected to characterize standalone performance will depend on the clinical task (clinical endpoint) and type of AI/ML output being evaluated. However, selection of performance metrics is not trivial, and a single measurement may not completely benchmark performance or fully capture the model task. As discussed previously, some common type of medical imaging AI/ML devices include segmentation, CADx, CADe, and computer-aided triage (CADt). Sokolova et al. provides an overview of the different evaluation metrics by dividing AI/ML into binary, multi-class, multi-labeled, and hierarchical tasks.[41] Other ways to categorize the metrics are based on whether the metric is prevalence dependent, applicable to binary truth, or applicable to multi-class truth.[42] Which measurements may be most appropriate depends on the task for which the AI/ML model was developed, the scale associated with the truth and the AI/ML output,[43] and which type of error should be most heavily weighted.[44]

2.4.2 *Subgroup analysis*

Subgroup assessment is an important component of standalone testing. In this case, model performance is assessed on individual or combined subgroups to better understand where the model may have performance limitations. Studies may need to be sized to yield a certain statistical precision of the performance on different subgroups, or studies may simply report the performance on different subgroups with accompanying confidence intervals. For the most part, this information is used in a regulatory submission to label the device, but subgroup performance may be critical in establishing substantial equivalence or safety and effectiveness when specific

performance claims are made for that subgroup. Depending on the clinical task and context, subgroup analyses found in AI/ML submissions are based on patient demographics, e.g., patient age, sex, race; image acquisition conditions, e.g., acquisition device and protocol; and disease type/presentation, e.g., disease subtype and lesion size/shape; or a combination of these characteristics.[31,45]

The subgroup analyses discussed above generally require prior knowledge of the important subgroups but some more automated techniques for identifying important subgroups, based on model audits[46] and schema completion[47] have been reported in the literature. Novel subgroup identification techniques may allow for generalizability analysis by automatically identifying important hidden stratifications impacting model performance.

2.4.3 *Repeatability and reproducibility*

A repeatability or reproducibility study refers to standalone assessments that investigate differences in AI/ML output from reimaging a patient with the same or different acquisition devices and conditions. Such studies are commonly included in *in vitro* diagnostic device submissions and may have value for submissions of AI/ML devices.[48–50] However, these designs have been less common in medical imaging applications, for example, when the study would have required exposing the patient to additional ionizing radiation. When appropriate, like scanning pathology slides multiple times with whole slide imaging systems or taking repeated pictures of a skin lesion with the camera of a mobile device, repeatability and reproducibility studies can demonstrate AI/ML device robustness and generalizability. More robust AI/ML devices will have higher repeatability/reproducibility across real-world use cases.

2.5 Multi-Reader Multi-Case Studies

AI/ML-enabled medical devices are often evaluated in the hands of clinicians especially when the intended use of the device is to assist clinicians in their clinical decision-making. Computer assistive AI/ML is particularly common in medical imaging applications. Here, we define a medical imaging reader study as a study in which readers, e.g., radiologists or pathologists, review and interpret medical images for a specified clinical task, e.g., diagnosis, and provide an objective interpretation, such as a rating of the likelihood that a condition is present. This is fundamentally different from a survey or questionnaire for the clinicians to indicate if they "like" some functionalities or features of an AI model, which is subjective and may not directly relate to how the AI/ML impacts clinician performance for a specific clinical task. When evaluating the clinical benefit of an AI/ML, it is ideal to have the conclusions of the clinical study generalize to both the intended patient population and the intended user population, i.e., readers. For this to occur, both readers and patient cases in the study should be representative of their respective populations, and both reader and case variability should be accounted for in the analysis method. The multi-reader multi-case (MRMC) study design is an approach that allows study conclusion to potentially generalize to both reader and patient populations when using the appropriate statistical analysis techniques.[51–53]

MRMC studies for medical imaging AI/ML typically consists of two arms: reading images without the AI/ML model and with the model for the clinical task for which the device is designed. This study allows reader performance with and without the AI/ML model to be compared. Many statistical methodologies for generalizing the performance to both the population of readers and the population of cases were developed including the Dorfman, Berbaum, and Metz (DBM) jackknife method,[54,55] the Obuchowski and Rockkett (OR) ANOVA model-based method,[56] the Beiden, Wagner, and Campbell bootstrap method,[57] and the Gallas U-statistics method.[53,58] While early developments of MRMC analysis methods focused on the area under the ROC curve (AUROC) as the preferred performance metric, most methods generalize to other endpoints, including binary outcome endpoints, e.g., the OR and U-statistics methods have been validated for binary outcome endpoints, such as sensitivity and specificity.[59,60] Many of these MRMC statistical methods have publicly available software tools, such as the updated OR method[61] and the U statistic method (iMRMC: software to do MRMC analysis of reader studies),[53] making MRMC assessment possible for non-statisticians.

The design of an MRMC reader study involves a number of considerations including patient data collection, establishment of a reference standard, recruitment and training of readers, and the study design along with other factors. Recent work has demonstrated statistical and practical tradeoffs to be considered when assigning cases to readers (fully crossed versus split-plot designs).[62,63] It is worth noting that MRMC studies for the assessment of imaging-based AI/ML are often retrospective and controlled "laboratory" studies, in which only the images and information related to the device of interest is presented to the readers, e.g., "image only" versus "image plus AI/ML output." This study approach is not fully consistent with the clinical reading scenario as physicians often have more information available, e.g., patient history, other clinical tests, and/or imaging exams. The studies are also often enriched with diseased cases when the natural prevalence of disease is low. The purpose of these design choices is to focus more directly on the AI/ML aid by limiting the impact of certain confounders and increasing the statistical power of the study rather than directly studying the "absolute" performance of clinicians in the real world.

While these specific design choices are often, but not always, acceptable in assessing the clinical performance of an AI/ML aid, efforts should be made to ensure the execution of the MRMC study is as close as possible to the clinical environment and identify/mitigate potential biases. For example, readers should be trained to appropriately use the AI/ML device, as this will help ensure the study design is not impacted as much by the readers learning on the fly how to use the AI/ML information effectively. In some cases, a prospective MRMC study may be necessary to clinically test an AI/ML device. For example, colonoscopy AI/ML aids have been assessed using a two-armed prospective MRMC study designs that more directly assesses the clinical impact of the AI/ML aid in device submissions.[64,65] It is also important to randomize cases, readers, and reading sessions to minimize bias. For more details on the design of MRMC studies, interested readers can refer to an FDA guidance document,[66] a consensus paper by Gallas et al.,[51] as well as a tutorial paper by Wagner et al.[52]

3 Discussion

There are many challenges to producing high quality medical imaging AI/ML that can effectively translate into clinical use. Saw et al. identified four overarching challenges for effectively implementing medical imaging AI/ML: data governance, algorithm robustness, stakeholder consensus, and legal liability.[67] While many of these challenges are intimately part of an AI/ML device review (data governance, algorithm robustness, and performance assessment), others, such as legal liability, generally fall outside of FDA's purview. Data governance concerns relate to developing effective policies and protocols for storing, securing, and maintaining data quality, including images, metadata, and reference standard (truth) labels.[67] Algorithm robustness concerns generally include how to reduce algorithmic bias and improve fairness across patients, groups, and sites. FDA is particularly concerned with how sponsor studies, often based on limited patient, group and site diversity, generalize to actual clinical practice across the United States. All AI/ML algorithms are biased to some extent because the data used to train the models are intrinsically a function of the population groups, disease conditions, clinical environments, and imaging technologies available during the collection process.[67] Therefore, robustness includes not only developing methods to measure and reduce important sources of bias but also defining fairness criteria for the AI/ML under appropriate operational conditions as well as integrating the appropriate level of transparency. Methodological assessment is also a critical aspect of the FDA review process. There are a large number of potential performance metrics, statistical approaches, and baseline comparators available for assessing AI/ML devices.[68] The challenge is to then determine the most meaningful data, metrics, methodological approaches, and success criteria for each unique AI/ML application while also controlling information leakage from the performance evaluation back into the device development process.

Many of these challenges are daunting and require contributions from developers, researchers, clinicians, patients, and regulatory communities. The agency's regulatory thinking and processes are evolving to address at least some of these challenges. One area of current interest is when should an AI/ML SaMD require a premarket submission for an algorithm change, including how to best regulate continuously learning AI/ML devices. FDA released a discussion paper

on this topic in 2019.[69] This discussion paper proposed a framework for regulating modifications to AI/ML-based SaMD that relies on the principle of a "predetermined change control plan" (PCCP). The discussion paper was a way to seek early input from groups and individuals outside the agency on this topic prior to development of any draft guidance document. FDA then released an AI/ML-based software as a medical device action plan in early 2021.[70] The action plan supports FDA's commitment to developing innovative approaches for regulating medical device software and other digital health technologies and was developed in direct response to the feedback received on the discussion paper. The FDA identified multiple challenges around regulating AI/ML and five main actions to pursue. These were[70]

(1) updating the proposed framework for AI/ML-based SaMD, including issuance of a draft guidance document on PCCPs,
(2) encouraging development and harmonization of good machine learning practices (GMLPs),
(3) holding a public workshop on medical device labeling to support transparency for users of AI/ML devices,
(4) supporting regulatory science efforts on the evaluation and improvement of AI/ML, and,
(5) advancing real-world performance pilots to provide additional clarity on what a real-world evidence generation program for AI/ML could look like.

FDA provides the device development community with guidance documents that represent FDA's current thinking and policies on regulatory issues (21 CFR 10.115(b)).[71] Guidance documents most often not only relate to the design, production, labeling, promotion, manufacturing, and testing of regulated products but can also discuss the processing, content, and evaluation, approval of submissions, or inspection and enforcement policies. The agency has two seminal guidance documents related to radiological imaging-based AI/ML devices. These documents discuss premarket notification [510(k)] submission details[72] and clinical performance assessment[66] of computer-assisted detection devices applied to radiology images and device data. FDA has a number of other potentially relevant guidance documents including guidance documents describing recommendations for performance data and software documentation for SaMD devices,[73] software validation,[74] and technical performance assessment for quantitative imaging devices.[75] In response to the 2021 AI/ML-based software as a medical device action plan, the FDA developed and recently released a draft guidance document entitled "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions."[13] This draft guidance document provides a framework for addressing some types of AI/ML algorithm modifications through a PCCP and is now available for review. FDA is encouraging individuals and organizations to submit comments for the FDA to consider before this guidance document is finalized.

To address the challenge around defining core GMLP concepts, FDA, Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency identified 10 guiding principles for the development of GMLP that help promote safe, effective, and high-quality AI/ML.[12] The agency is also participating in collaborative communities,[76] such as the AFDO/RAPS healthcare products collaborative community, which is focused on GMLP development, and the pathology innovation collaborative community, where AI/ML is a major topic driving community activities.

Many groups have recognized a lack of transparency in AI/ML-enabled medical devices. For example, van Leeuwen et al.[77] determined that only 36/100 identified European CE-marked AI/ML products had peer-reviewed performance information available with only 18/36 demonstrating at least some potential clinical impact. FDA makes device summaries available for all cleared/approved medical devices, including medical imaging AI/ML, but these summaries may lack some types of information needed by specific stakeholders. Transparency is not only a concern for regulators but across the AI/ML landscape. One example of this is a lack of code and data sharing hindering the ability of outside groups to perform reproducibility studies and assessments for many AI/ML algorithms appearing in the peer-reviewed literature.[78] Therefore, there remains transparency challenges including determining exactly what information is needed, who needs it and when, and how best to present this information to the audience. The FDA held a virtual

public workshop on the transparency of AI/ML-enabled medical devices in October, 2021, to hear from stakeholders on considerations for achieving transparency for users of AI/ML-enabled medical devices, and to gather input on the types of information that would be helpful for a manufacturer to include in device labeling and public facing documents to support transparency and potentially AI/ML explainability.[79]

Many AI/ML performance assessment and algorithm robustness challenges still depend on the development of novel approaches and tools. The Office of Science and Engineering Labs (OSEL) within CDRH is focused on advancing AI/ML regulatory science.[14] Regulatory science is the science of developing new tools, standards, and approaches to assess the safety, efficacy, quality, and performance of FDA-regulated products. OSEL has an AL/ML regulatory science program conducting research to ensure patients have access to safe and effective medical AI/ML.[80] This research is addressing major scientific gaps and challenges including (a) a lack of methods for the enhancement of AI/ML model training for clinical datasets that are typically much smaller than nonclinical datasets, (b) a lack of clear definitions or understanding of artifacts, limitations, and failure modes for deep-learning models in the denoising and reconstruction of medical images, (c) a lack of a clear reference standard for assessing the accuracy of AI/ML-based quantitative imaging and radiomics tools, (d) lack of assessment techniques to evaluate the trustworthiness of adaptive and autonomous AI/ML devices, e.g., continuously learning models, and (e) a lack of systematic approaches to address the robustness of various AI/ML input factors, such as data acquisition factors, patient demographics, and disease factors, to patient outcomes in a regulatory submission.

The AI/ML research program[80] strives to address these challenges by conducting peer-reviewed research to develop and understand methods for enhanced AI/ML training, developing systematic approaches for understanding AI/ML robustness, and assessing novel test methodologies to evaluate fixed and continuously learning AI/ML performance in both the premarket and real-world settings, to name just a few areas of ongoing research.

Some of the regulatory science projects being conducted as part of the OSEL AI/ML research program include a recent investigation developing a cooperative labeling technique to incorporate weakly labeled data into the training of a deep learning AI/ML model for lung nodule detection in CT.[81] This study showed that the inclusion of weakly labeled data leads to a 5% improvement in lung nodule detection performance when the number of expert annotations is limited. Another approach for addressing small dataset sizes is to augment available data with synthetic datasets. Cha et al.[82] compared detection performance when the network was trained using different percentages of real and synthetic mammograms. Synthetic mammograms were generated using *in silico* breast and lesion models followed by the creation of synthetic mammograms of the breast models. The results showed that a statistically significant improvement in detection sensitivity can be achieved when synthetic images are added to real mammograms in algorithm training. These results indicate that novel approaches for augmenting and expanding training data, e.g., using generative, *in silico* or phantom-based augmentation, could play a role in reducing burden and improving the performance and robustness of medical imaging AI/ML devices.

OSEL is also exploring approaches for efficiently and effectively utilizing limited data in AI/ML algorithm modifications. A Centers of Excellence in Regulatory Science and Innovation (CERSI) collaboration between the University of California at San Francisco and OSEL investigated whether an online logistic recalibration and revision procedure can be designed with performance guarantees on updates to an original "static" AI/ML model. The overall goal was to avoid the risk of deteriorating model performance that may inadvertently result from an AI/ML model update. The team designed two procedures for continual recalibration or revision of an underlying AI/ML model. These procedures guarantee that the updated models are noninferior to the static model and often produce model revisions that improve over time as is desired.[83] Result from an empirical evaluation via simulations and a real-world study predicting chronic obstructive pulmonary disease risk showed that both methods outperformed the static model and other common online revision techniques.

Another OSEL regulatory science effort is examining the expected time saving from a CADt device implemented within a clinical setting. Ideally, a CADt device prioritizes patients with a time sensitive condition so these patients are evaluated more quickly. However, quantifying the time savings is challenging because of the complex and heterogeneous clinical environments a

CADt device may be used in. OSEL scientists have developed a theoretical method, based on queueing theory, to quantify the wait-time-savings of CADt in various clinical settings.[84] The theoretical model was validated via simulation studies and allows model users to investigate CADt performance under various clinical settings, including changes in disease prevalence, patient arrival rate, radiologist reading rate, number of radiologists on-site, and the presence of emergency patient images.

OSEL is also developing statistical methods for assessing AI/ML device performance. For example, an agreement endpoint may be the most acceptable metric for assessment of AI/ML devices that output a quantitative measurement derived from a medical image, especially when the reference method is clinicians estimating the same value. A recent OSEL paper reports on a three-way mixed effects ANOVA technique for estimating MRMC agreement in a statistically rigorous manner.[85] Another recent study investigated a method for controlling "information leakage" through, for example, the repeated reuse of test data in AI/ML device evaluation studies. Test data reuse is related to the problem of privacy loss due to repeated queries of information from a database. Differential privacy methods have been developed to address the latter problem and have also been applied to the former test data reuse problem as well. This OSEL study[86] extended the reusable holdout mechanism of Dwork et al.[87] to the more common AUROC endpoint used in AI/ML device assessment and showed that this method substantially reduced overfitting to the test data, even when the test dataset is small, but comes at the cost of increased uncertainty in the reported performance.

To accelerate the transfer of regulatory science methods into the fast-evolving AI/ML technological landscape, outcomes from these research efforts are being released as regulatory science tools (RSTs).[88] RSTs are peer-reviewed computational or physical phantoms, methods, datasets, computational models, and simulation pipelines designed to support the assessment of safety or effectiveness of a medical device or emerging technology. These tools are well characterized for their applications and are made broadly available through a CDRH/OSEL public catalog of more than 100 RSTs, including AI/ML assessment tools.[88] One available AI/ML assessment RST is the iMRMC tool that can be used to assist investigators in analyzing and sizing MRMC reader studies.[53] This catalog is being expanded as new tools are developed.

4 Conclusion

This paper discussed FDA medical device review processes, including device types, product classifications, and regulatory pathways for medical imaging AI/ML devices. The device class (classes I, II, or III) and the regulatory pathway (PMA, 510k, or De Novo) are based on the level of risk associated with an AI/ML device and informed by both the technological characteristics and intended use of the device. Over five hundred medical devices incorporating AI/ML technology have been granted marketing authorization by the FDA through a combination of the PMA, 510k, and De Novo regulatory pathways, with the majority of these devices analyzing radiological image data. Even though the history of AI/ML devices on the market is long, AI/ML is still a fast-changing and evolving technology that presents novel regulatory challenges for developing robust assessment methods and providing effective regulatory oversight. To solve these challenges, FDA is conducting and facilitating AI/ML regulatory science research that focuses on allowing innovation to flourish through least burdensome regulatory methods while still assuring that patients have timely and continued access to safe, effective, and high-quality AI/ML devices.

Disclosures

The authors have no financial conflicts of interest to disclose.

Disclaimers

The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

References

1. S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database," *NPJ Digit. Med.* **3**(1), 118 (2020).
2. U.S. Food and Drug Administration, "PAPNET testing system: summary of safety and effectiveness," FDA webpage (1995), https://www.accessdata.fda.gov/cdrh_docs/pdf/p940029.pdf (accessed 16 November 2022).
3. U.S. Food and Drug Administration, "Requests for feedback and meetings for medical device submissions: the Q-submission program," (2019), https://www.fda.gov/media/114034/download (issued 6 Januray 2021).
4. "IMDRF: software as a medical device (SaMD): key definitions," (2013), http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf (accessed 9 January 2023).
5. U.S. Food and Drug Administration, "Artificial intelligence and machine learning (AI/ML)-enabled medical devices," FDA webpage (2022), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (accessed 16 November 2022).
6. U.S. Food and Drug Administration, "OMJ: chest x-ray computer aided detection," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=5665 (accessed 16 November 2022).
7. U.S. Food and Drug Administration, "QDQ: radiological computer assisted detection/diagnosis software for lesions suspicious for cancer," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfTPLC/tplc.cfm?id=5679 (accessed 16 November 2022).
8. U.S. Food and Drug Administration, "QAS: radiological computer-assisted triage and notification software," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=5677 (accessed 16 November 2022).
9. U.S. Food and Drug Administration, "QJU: image acquisition and/or optimization guided by artificial intelligence," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=5685 (accessed 16 November 2022).
10. U.S. Food and Drug Administration, "QNP: gastrointesinal lesion software detection system," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=2257 (accessed 16 November 2022).
11. U.S. Food and Drug Administration, "QPN: software algorithm device to assist users in digital pathology," FDA product classification (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpcd/classification.cfm?id=5275 (accessed 16 November 2022).
12. U.S. Food and Drug Administration, "Good machine learning practice for medical device development: guiding principles," FDA webpage (2021), https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (accessed 13 January 2023).
13. U.S. Food and Drug Administration, "Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions," U.S. Food and Drug Administration, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial (2023).
14. B. D. Gallas et al., "FDA fosters innovative approaches in research, resources and collaboration," *Nat. Mach. Intell.* **4**(2), 97–98 (2022).
15. K. Lekadir et al., "Future-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging," arXiv:2109.09658 (2021).
16. L. Hadjiiski et al., "AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging," *Med. Phys.* **50**(2), e1–e24 (2023).
17. T. J. Bradshaw et al., "Nuclear medicine and artificial intelligence: best practices for algorithm development," *J. Nucl. Med.* **63**(4), 500–510 (2022).
18. E. Abels et al., "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association," *J. Pathol.* **249**(3), 286–294 (2019).
19. U.S. Food and Drug Administration, "The 510(k) program: evaluating substantial equivalence in premarket notifications [510(k)]: guidance for industry and Food and Drug Administration staff," (2014), https://www.fda.gov/media/82395/download (accessed 31 October 2021).
20. U.S. Food and Drug Administration, "General controls for medical devices," FDA webpage (2018), https://www.fda.gov/medical-devices/regulatory-controls/general-controls-medical-devices (accessed 13 January 2023).
21. U.S. Food and Drug Administration, "Regulatory controls," FDA webpage (2018), https://www.fda.gov/medical-devices/overview-device-regulation/regulatory-controls (accessed 16 November 2022).
22. U.S. Food and Drug Administration, "Product classification database," FDA webpage (2023), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/PCDSimpleSearch.cfm (accessed 9 January 2023).

23. U.S. Food and Drug Administration, "De Novo classification process (evaluation of automatic class III designation) - guidance for industry and FDA staff," (2021), https://www.fda.gov/media/72674/download (issued 5 October 2021).
24. U.S. Food and Drug Administration, "GI genius: FDA reclassification order," (2021), https://www.accessdata.fda.gov/cdrh_docs/pdf20/DEN200055.pdf (accessed 16 November 2022).
25. U.S. Food and Drug Administration, "QauntX: FDA reclassification order," (2020), https://www.accessdata.fda.gov/cdrh_docs/pdf17/DEN170022.pdf (accessed 9 January 2023).
26. U.S. Food and Drug Administration, "Paige prostate: FDA reclassification order," (2021), https://www.accessdata.fda.gov/cdrh_docs/pdf20/DEN200080.pdf (accessed 15 November 2022).
27. F. S. Abas et al., "Computer-assisted quantification of CD^{3+}T cells in follicular lymphoma," *Cytometry Part A* **91A**(6), 609–621 (2017).
28. N. Petrick et al., "Evaluation of computer-aided detection and diagnosis systems," *Med. Phys.* **40**, 087001–087017 (2013).
29. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (corrected 12th printing), Springer, New York (2017).
30. W. Chen et al., "Chapter 23: a regulatory science perspective on performance assessment of machine learning algorithms in imaging," in *Machine Learning for Brain Disorders*, O. Colliot, Ed., Springer (2023).
31. V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Am. Med. Assoc.* **316**(22), 2402–2410 (2016).
32. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York (1990).
33. A. C. Justice, K. E. Covinsky, and J. A. Berlin, "Assessing the generalizability of prognostic information," *Ann. Intern. Med.* **130**(6), 515–524 (1999).
34. K. G. M. Moons et al., "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration," *Ann. Intern. Med.* **162**(1), W1–W73 (2015).
35. B. Du et al., "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.* **47**(1), 14–26 (2017).
36. S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1936–1949 (2014).
37. U.S. Food and Drug Administration, "The least burdensome provisions: concept and principles - guidance for industry and FDA staff," (2019), https://www.fda.gov/media/73188/download (issued 5 February 2019).
38. U.S. Food and Drug Administration, "Recommended content and format of non-clinical bench performance testing information in premarket submissions - guidance for industry and Food and Drug Administration staff," 2019, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/recommended-content-and-format-non-clinical-bench-performance-testing-information-premarket (accessed 26 April 2019).
39. E. W. Steyerberg, "Overfitting and optimism in prediction models," in *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, E. W. Steyerberg, Ed., pp. 95–112, Springer, Cham, Switzerland (2019).
40. A. Gossmann et al., "Considerations in the assessment of machine learning algorithm performance for medical imaging," in *Deep Learning for Medical Image Analysis*, K. Zhou and H. Greenspan and D. Shen, eds., Elsevier Inc. (2022).
41. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* **45**(4), 427–437 (2009).
42. C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.* **30**(1), 27–38 (2009).
43. S. S. Stevens, "On the theory of scales of measurement," *Science* **103**(2684), 677–680 (1946).
44. A. Reinke et al., "Common limitations of image processing metrics: a picture story," arXiv:2104.05642 (2021).
45. P. Wang et al., "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study," *Gut* **68**(10), 1813–1819 (2019).
46. V. Mahajan et al., "The algorithmic audit: working with vendors to validate radiology-AI algorithms: how we do it," *Acad. Radiol.* **27**(1), 132–135 (2020).
47. L. Oakden-Rayner et al., "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," (2019).
48. U.S. Food and Drug Administration, "Establishing the performance characteristics of in vitro diagnostic devices for the detection or detection and differentiation of influenza viruses: guidance for industry and FDA staff," FDA guidance document (2011), https://www.fda.gov/media/71519/download (accessed 15 November 2022).
49. ISO 5725-2 Writing Committee, "Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method," (ISO 5725-2), International Organization for Standardization, https://www.iso.org/standard/69419.html (2019).

50. R. McEnroe et al., "Evaluation of precision of quantitative measurement procedure; approved guideline, (EPO5-A3)," Clinical Laboratory Standards Institute, https://clsi.org/media/1438/ep05a3_sample.pdf (2014).
51. B. D. Gallas et al., "Evaluating imaging and computer-aided detection and diagnosis devices at the FDA," *Acad. Radiol.* **19**(4), 463–477 (2012).
52. R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: a tutorial review," *Acad. Radiol.* **14**(6), 723–748 (2007).
53. U.S. Food and Drug Administration, "iMRMC: software for the statistical analysis of multi-reader multi-case studies," FDA webpage (2022), https://www.fda.gov/medical-devices/science-and-research-medical-devices/imrmc-software-statistical-analysis-multi-reader-multi-case-reader-studies (accessed 27 January 2023).
54. D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: generalization to the population of readers and cases with the jackknife method," *Invest. Radiol.* **27**, 723–731 (1992).
55. L. Hadjiiski et al., "Breast masses: computer-aided diagnosis with serial mammograms," *Radiology* **240**(2), 343–356 (2006).
56. N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an ANOVA approach with dependent observations," *Commun. Stat. Simul. Comput.* **24**(2), 285–308 (1995).
57. S. V. Beiden, R. F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis," *Acad. Radiol.* **7**(5), 341–349 (2000).
58. B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**(3), 353–362 (2006).
59. B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am. A* **24**(12), B70–B80 (2007).
60. W. Chen et al., "Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing," *JMIOBU* **1**(3), 031011–031011 (2014).
61. S. Hillis, K. Schartz, and M. Madsen, "Medical Image Perception Laboratory, Department of Radiology: Software," https://perception.lab.uiowa.edu/software-0 (accessed 13 June 2023).
62. N. A. Obuchowski, B. D. Gallas, and S. L. Hillis, "Multi-reader ROC studies with split-plot designs: a comparison of statistical methods," *Acad. Radiol.* **19**(12), 1508–1517 (2012).
63. W. Chen, Q. Gong, and B. Gallas, "Efficiency gain of paired split-plot designs in MRMC ROC studies," *Proc. SPIE* **10577**, 10577F (2018).
64. A. Repici et al., "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial," *Gastroenterology* **159**(2), 512–520.e7 (2020).
65. U.S. Food and Drug Administration, "GI genius: FDA decision summary," FDA decision summary (2021), https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN200055.pdf (accessed 10 January 2023).
66. U.S. Food and Drug Administration, "Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification [510(k)] submissions: guidance for industry and Food and Drug Administration staff," FDA guidance document (2020), https://www.fda.gov/media/77642/download (issued 22 January 2020).
67. S. N. Saw and K. H. Ng, "Current challenges of implementing artificial intelligence in medical imaging," *Phys. Med.* **100**, 12–17 (2022).
68. G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *NPJ Digit. Med.* **5**(1), 48 (2022).
69. U.S. Food and Drug Administration, "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) - discussion paper and request for feedback," https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (accessed 31 October 2021).
70. U.S. Food and Drug Administration, "Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan," https://www.fda.gov/media/145022/download (accessed 22 October 2021).
71. U.S. Food and Drug Administration, "FDA guidances," (2022), https://www.fda.gov/industry/fda-basics-industry/guidances (accessed 18 May 2022).
72. U.S. Food and Drug Administration, "Computer-assisted detection devices applied to radiology images and radiology device data: premarket notification [510(k)] submissions," FDA guidance document (2022), https://www.fda.gov/media/77635/download (issued 28 September 2022).
73. U.S. Food and Drug Administration, "Guidance for the content of premarket submissions for software contained in medical devices - guidance for industry and FDA staff," (2005), https://www.fda.gov/media/73065/download (issued 11 May 2005).
74. U.S. Food and Drug Administration, "General principles of software validation - guidance for industry and FDA staff," (2002), https://www.fda.gov/media/73141/download (issued 11 January 2002).

75. U.S. Food and Drug Administration, "Technical performance assessment of quantitative imaging in radiological device premarket submissions - guidance for industry and FDA staff," (2022), https://www.fda.gov/media/123271/download (issued 16 June 2022).
76. U.S. Food and Drug Administration, "Collaborative communities: addressing health care challenges together," FDA webpage (2021), https://www.fda.gov/about-fda/cdrh-strategic-priorities-and-updates/collaborative-communities-addressing-health-care-challenges-together (accessed 27 January 2023).
77. K. G. van Leeuwen et al., "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence," *Eur. Radiol.* **31**(6), 3797–3804 (2021).
78. B. Haibe-Kains et al., "Transparency and reproducibility in artificial intelligence," *Nature* **586**(7829), E14–E16 (2020).
79. U.S. Food and Drug Administration, "Virtual public workshop - transparency of artificial intelligence/machine learning-enabled medical devices," FDA webpage (2021), https://www.fda.gov/medical-devices/workshops-conferences-medical-devices/virtual-public-workshop-transparency-artificial-intelligencemachine-learning-enabled-medical-devices (accessed 20 January 2023).
80. U.S. Food and Drug Administration, "Artificial intelligence and machine learning program: research on AI/ML-based medical devices," FDA webpage (2021), https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-and-machine-learning-program-research-aiml-based-medical-devices (accessed 13 January 2023).
81. M. Maynord et al., "Semi-supervised training using cooperative labeling of weakly annotated data for nodule detection in chest CT," *Med. Phys.* (2023).
82. K. Cha et al., "Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning," *JMIOBU* **7**(1), 012703 (2019).
83. J. Feng et al., "Bayesian logistic regression for online recalibration and revision of risk prediction models with performance guarantees," *J. Am. Med. Inf. Assoc.* **29**(5), 841–852 (2022).
84. Y. L. E. Thompson et al., "Wait-time-saving analysis and clinical effectiveness of computer-aided triage and notification (CADt) devices based on queueing theory," *Proc. SPIE* **12035**, 120350Q (2022).
85. S. Wen and B. D. Gallas, "Three-way mixed effect ANOVA to estimate MRMC limits of agreement," *Stat. Biopharm. Res.* **14**(4), 532–541 (2022).
86. A. Gossmann et al., "Test data reuse for the evaluation of continuously evolving classification algorithms using the area under the receiver operating characteristic curve," *SIAM J. Math. Data Sci.* **3**(2), 692–714 (2021).
87. C. Dwork et al., "The reusable holdout: preserving validity in adaptive data analysis," *Science* **349**(6248), 636–638 (2015).
88. U.S. Food and Drug Administration, "Catalog of regulatory science tools to help assess new medical devices," (2020), https://www.fda.gov/medical-devices/science-and-research-medical-devices/catalog-regulatory-science-tools-help-assess-new-medical-devices (accessed 13 November 2020).

Nicholas Petrick, PhD, is a deputy director for the Division of Imaging, Diagnostics, and Software Reliability at the Center for Devices and Radiological Health, U.S. Food and Drug Administration. He received his PhD from the University of Michigan in Electrical Engineering-Systems and is a fellow of the American Institute of Medical and Biomedical Engineering and SPIE. His research focuses on medical artificial intelligence and assessment methods for a range of medical imaging hardware and artificial intelligence tools.

Weijie Chen, PhD, is a research physicist in the Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories at the Center for Devices and Radiological Health, U.S. Food and Drug Administration. He received his PhD from the University of Chicago in Medical Physics and is a fellow of SPIE. His research focuses on assessment methods for artificial intelligence and machine learning devices in radiology and pathology applications and statistical methods in medical imaging.

Jana G. Delfino, PhD, is a deputy director for the Division of Imaging, Diagnostics, and Software Reliability in the Center for Devices and Radiological Health at the U.S. Food and Drug Administration. She received a BS degree in agricultural engineering from the University of California at Davis and a PhD in biomedical engineering from the Georgia Institute of Technology and Emory University. She is active in developing guidance regarding the incorporation of artificial intelligence into radiological devices.

Brandon D. Gallas, PhD, provides mathematical, statistical, and modeling expertise to the evaluation of medical imaging devices at the FDA. His main areas of research are image quality,

computer-aided diagnosis, imaging physics, and the design, execution, and statistical analysis of reader studies (https://www.fda.gov/medical-devices/science-and-research-medical-devices/imrmc-software-statistical-analysis-multi-reader-multi-case-reader-studies). Recently, he has been investigating pathologist performance and agreement using whole slide imaging devices and the microscope in an effort to create a dataset of pathologist annotations to validate AI/ML (https://github.com/DIDSR/HTT).

Yanna Kang, PhD, is an assistant director for the Mammography and Ultrasound Team at the Center for Devices and Radiological Health, U.S. Food and Drug Administration. She received her BS degree from Northeastern University, China, in computer science and her PhD from the University of Pittsburgh in Intelligent Systems, a multidisciplinary graduate program dedicated to applied artificial intelligence. Her research interests include applied artificial intelligence in medical imaging and clinical text data.

Daniel Krainak, PhD, is the assistant director for the magnetic resonance and nuclear medicine devices team at the Center for Devices and Radiological Health, U.S. Food and Drug Administration (FDA). He joined the FDA in 2011 after completing his PhD in biomedical engineering at Northwestern University. He participates in the FDA review of radiological devices, imaging biomarkers, and radiological imaging in therapeutic medical product clinical trials.

Berkman Sahiner, PhD, received his PhD in electrical engineering and computer science from the University of Michigan, Ann Arbor. He is a senior biomedical research scientist in the Division of Imaging of Diagnostics and Software Reliability at the Center for Devices and Radiological Health of the U.S. FDA. His research is focused on the evaluation of medical imaging and computer-assisted diagnosis devices, including devices that incorporate machine learning and artificial intelligence. He is a fellow of SPIE and AIMBE.

Ravi K. Samala, PhD, is a regulatory scientist in the Division of Imaging, Diagnostics, and Software Reliability, in the Center for Devices and Radiological Health at the U.S. Food and Drug Administration with an adjunct appointment in the Department of Radiology, University of Michigan. He has extensive experience in machine learning within a wide range of medical imaging modalities and conducted research in academic and regulatory science.