Chapter 23

A Regulatory Science Perspective on Performance Assessment of Machine Learning Algorithms in Imaging

Weijie Chen, Daniel Krainak, Berkman Sahiner, and Nicholas Petrick

Abstract

This chapter presents a regulatory science perspective on the assessment of machine learning algorithms in diagnostic imaging applications. Most of the topics are generally applicable to many medical imaging applications, while brain disease-specific examples are provided when possible. The chapter begins with an overview of US FDA's regulatory framework followed by assessment methodologies related to ML devices in medical imaging. Rationale, methods, and issues are discussed for the study design and data collection, the algorithm documentation, and the reference standard. Finally, study design and statistical analysis methods are overviewed for the assessment of standalone performance of ML algorithms as well as their impact on clinicians (i.e., reader studies). We believe that assessment methodologies and regulatory science play a critical role in fully realizing the great potential of ML in medical imaging, in facilitating ML device innovation, and in accelerating the translation of these technologies from bench to bedside to the benefit of patients.

Key words Machine learning, Performance assessment, Standalone performance, Reader study, Statistical analysis plan, Regulatory science

1 Introduction

Machine learning (ML) technologies are being developed at an ever-increasing pace in a variety of medical imaging applications [1]. Particularly in brain imaging, the past decade has witnessed a spectacular growth of ML development for the diagnosis, prognosis, and treatment of brain disorders [2]. One of the ultimate goals of these developments is to translate safe and effective technologies to the clinic to benefit patients. Regulatory oversight plays a key role in this translation. The mission of the Center for Devices and Radiological Health (CDRH) at the US Food and Drug Administration (US FDA) is to "assure that patients and providers have timely and continued access to safe, effective, and high-quality

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_23,

medical devices."[1] This chapter discusses performance assessment of machine learning algorithms in imaging applications from a regulatory science perspective. Regulatory science is the science of developing new tools, standards, and approaches to assess the safety, efficacy, quality, and performance of all FDA-regulated products.[2]

We begin with clarifications of the scope of this chapter. First, following an overview of the US FDA's regulatory framework for medical imaging and related ML devices, the primary topics we discuss are about concepts, basic principles, and methods for performance assessment of ML algorithms in the arena of *regulatory science* but not *regulatory policy*. As such, these topics are not necessarily relevant to every regulatory submission. The question of which components should be included in a specific regulatory submission is a regulatory decision depending on factors such as the risk of the device, impact on clinical practice, complexity of the technology, precedents, and so on and is beyond the scope of this chapter. Second, the topics are *selected* based on our experience and expertise but are not intended to be comprehensive. For example, software engineering and cybersecurity are important aspects of ML devices but are beyond the scope of this chapter. Third, as discussed in earlier chapters of this book, ML algorithms are developed for both *imaging* and *non-imaging* modalities for treating brain disorders. We focus on imaging applications. Moreover, while this book is on brain disorders, most of the discussions in this chapter are applicable to ML algorithms in general imaging applications unless noted otherwise. Lastly, while the assessment methods are well established to the best of our knowledge at the time of writing, we acknowledge that ML techniques and assessment methodologies are active areas of research and better methods may become available and adopted by researchers, developers, and regulatory agencies alike in the future. To give the readers a more specific sense of the scope of applications that are relevant to our discussions, we reviewed, via the American College of Radiology (ACR) and FDA public databases, some ML devices for brain disorders that were authorized by the FDA in recent years and summarized major scope characteristics including the imaging modalities, functionalities, and types of ML algorithms (*see* Table 1).

The rest of the chapter begins with an overview of US FDA's regulatory framework followed by topics on assessment methodologies related to ML devices in medical imaging. Rationale, methods, and issues are discussed for study design and data collection

[1] https://www.fda.gov/about-fda/center-devices-and-radiological-health/cdrh-mission-vision-and-shared-values

[2] https://www.fda.gov/science-research/science-and-research-special-topics/advancing-regulatory-science

Table 1
Summary characteristics of exemplar FDA-cleared ML devices for brain disorders

Modality	CT (contrast or non-contrast), CTA, MRI, PET, SPECT
Functionality	Triage and notification (e.g., for intracranial hemorrhage); segmentation, quantification, and feature measurements; analysis and visualization; computer-aided diagnosis; denoising, enhancement; auto-contouring/segmentation of organs at risk or tumors for radiation therapy of head and neck tumors
ML algorithms	Hand-crafted feature extraction and computerized classifiers; deep learning neural networks

CT computed tomography, *CTA* computed tomography angiography, *MRI* magnetic resonance imaging, *PET* positron emission tomography, *SPECT* single photon emission computed tomography. Summary based on a sampled review of public databases at ACR (https://models.acrdsi.org/) and FDA (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices) websites. The table aims to give a general overview of the scope of devices available. For specific devices that work for a certain imaging modality with certain functionalities, please refer to the cited databases

(Subheading 3), algorithm documentation (Subheading 4), and reference standard (Subheading 5). Finally, performance assessment methodologies are overviewed including the standalone performance assessment of ML algorithms (Subheading 6), assessment of ML algorithms in the hands of clinicians (i.e., reader studies; Subheading 7), and general considerations for the statistical analysis (Subheading 8). The relationships among these topics are illustrated in Fig. 1. Performance assessment of ML devices is necessary in both premarket and postmarket environments. Premarket studies are for the assessment of safety and effectiveness before the device is authorized for marketing by a regulatory body. Some premarket studies are used in the context of device development to refine and iterate on device design. Other premarket studies are intended for review by regulatory bodies to help assess the safety and effectiveness prior to marketing authorization. Postmarket studies are for clinical use and epidemiology, maintenance, and modifications. The selected topics to be discussed in this chapter belong to premarket performance assessment.

2 Regulatory Framework

CDRH Learn[3] provides readers an excellent resource to better understand overall medical device regulation.

2.1 Overview

The US FDA classifies medical devices into three classes, Classes I, II, and III. The classification determines the extent of regulatory controls necessary to provide reasonable assurance of the safety and

[3] https://www.fda.gov/training-and-continuing-education/cdrh-learn

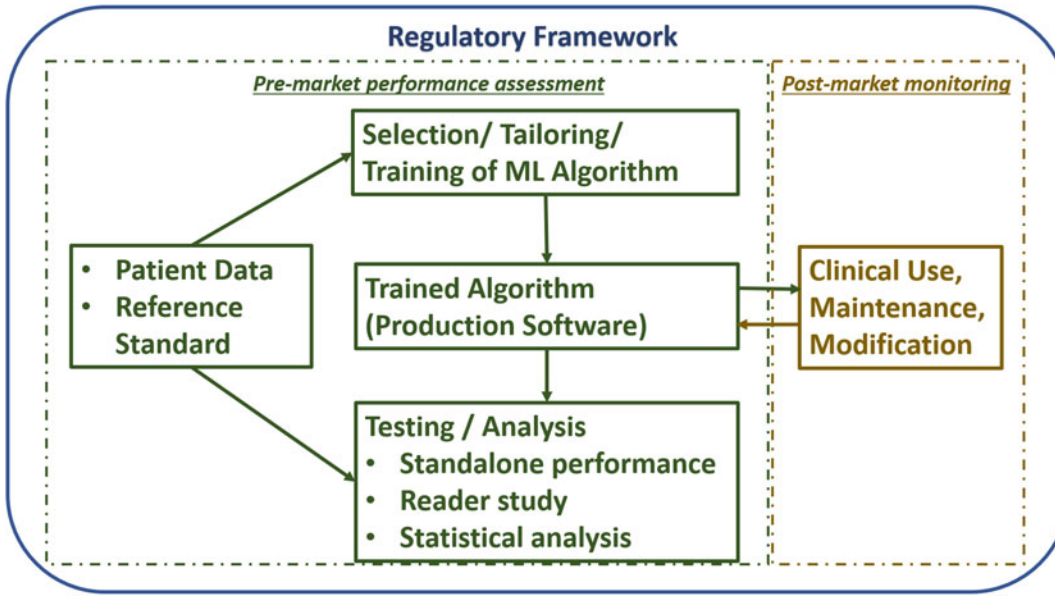

Fig. 1 ML performance assessment methods in the context of US FDA's regulatory framework

effectiveness of the device. The device classification tends to increase with increasing degree of risk, and the appropriate types of controls applicable to the device depend on the device classification. There are three types of regulatory controls: general controls, special controls, and premarket authorization requirements. General controls include the basic provisions applicable to medical devices of the Food, Drug, and Cosmetic Act and apply to all medical devices. They include provisions that relate to adulteration; misbranding; device registration and listing; premarket notification; banned devices; notification, including repair, replacement, or refund; records and reports; restricted devices; and good manufacturing practices.[4] Special controls apply to Class II devices and are published in the Code of Federal Regulations under the specific device type. Some examples of special controls include labeling, testing, design specifications, software life cycle documentation activities, and usability assessments.

The US FDA requirements for premarket submissions differ between the device classes. To receive FDA approval, sponsors of Class III devices, generally considered the highest risk devices, must demonstrate a reasonable assurance of safety and effectiveness. Sponsors of Class I and II device must demonstrate substantial equivalence between their new device and a legally marketed device through the premarket notification process (i.e., the 510 [k] Program), unless the product class is exempt from premarket

[4] https://www.fda.gov/medical-devices/regulatory-controls/general-controls-medical-devices

notification. Substantial equivalence is a comparative analysis that includes a comparison of the intended use, technological characteristics, and performance testing. For device classifications that include defined special controls (generally published in the Code of Federal Regulations or in an order granting a request for reclassification), the sponsor must also demonstrate that they have fulfilled all the necessary special controls as part of the premarket notification process and to avoid marketing an adulterated or misbranded device.

The De Novo classification process is a pathway to 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 [3]. Devices of a new type that FDA has not previously classified are "automatically" or "statutorily" classified into Class III by the FD&C Act, regardless of the level of risk they pose or the ability of general and special controls to assure safety and effectiveness. Section 513(f)(2) of the FD&C Act allows manufacturers to submit a De Novo request to FDA for devices "automatically" classified into Class III by operation of Section 513(f)(1). In essence, a De Novo is a request for classification for a novel device that would otherwise be classified as a Class III device. During review of a De Novo request, the FDA evaluates whether general controls or general and special controls are adequate to provide a reasonable assurance of safety and effectiveness for the identified classification of the device.

FDA regulates products based upon the device characteristics (e.g., what is it? what does it do?) and the intended use of the device. The submission type and performance data necessary to obtain marketing authorization depends on the device classification, technological characteristics, and intended use. Understanding the technological characteristics is often a more straightforward exercise compared to the determination of the intended use of the product when attempting to determine the appropriate regulatory pathway and necessary supporting data. Intended use means the general purpose of the device or its function and encompasses the indications for use [4]. The indications for use, as defined in 21 CFR 814.20(b)(3)(i), describes 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. The intended use of a device is one criterion that determines whether a device can be cleared for marketing through the 510(k) process or must be evaluated as a Class III device (premarket approval) or, if appropriate, a De Novo request. Section 513(i)(1)(E)(i) of the FD&C Act provides that the FDA's determination of intended use of a device "shall be based upon the proposed labeling." A device may have a variety of different indications for use and

intended uses (e.g., output a measurement for users, identify patients eligible for a particular treatment, estimate prognostic cancer risk, or predict a patient's response to therapy). The data needed to support these different intended uses and indications are different.

2.2 Imaging Device Regulation

A majority of medical image processing devices have been classified as Class II devices. Most of the software-only devices or software as a medical device that are intended for image processing have been classified under 21 CFR 892.2050 as picture archiving and communications systems. On April 19, 2021, FDA updated the name of the regulation 21 CFR 892.2050 to "medical image management and processing system." There are no published, mandatory specific special controls related to software-only devices classified under 21 CFR 892.2050, and therefore, the primary resource to understand the legal requirements for performance data associated with these devices is the comparative standard of substantial equivalence as described in detail in the guidance document on the 510 (k) Program [4]. In contrast, several devices more recently classified under the De Novo pathway have specific special controls that manufacturers marketing such devices must adhere to.

Devices originally classified via the De Novo pathway often include special controls defined in the CFR describing requirements for manufacturers of these devices. Devices that may implement machine learning that include software or software-only devices must adhere to the special controls defined in the specific regulations associated with the appropriate device class. The classification with the associated special controls is published with a Federal Register notice and appears in the Electronic Code of Federal Regulations (eCFR).[5] A De Novo classification, including any special control, is effective on the date the order letter is issued granting the De Novo request [3]. For the specific examples cited below, the De Novo submission (DEN number) is cited for classifications that have not been published in CFR at the time of writing, and the associated order with special controls may be found by searching FDA's De Novo database.[6] Examples include:

- 21 CFR 870.2785 (DEN200019): Software for optical camera-based measurement of pulse rate, heart rate, breathing rate, and/or respiratory rate
- 21 CFR 870.2790 (DEN200038): Hardware and software for optical camera-based measurement of pulse rate, heart rate, breathing rate, and/or respiratory rate

[5] https://www.ecfr.gov/cgi-bin/ECFR?page=browse

[6] https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm

- 21 CFR 876.1520 (DEN200055): Gastrointestinal lesion software detection system
- 21 CFR 892.2060 (DEN170022): Radiological computer-assisted diagnostic software for lesions suspicious of cancer
- 21 CFR 892.2070: Medical image analyzer
- 21 CFR 892.2080 (DEN170073): Radiological computer-aided triage and notification software
- 21 CFR 892.2090 (DEN180005): Radiological computer-assisted detection and diagnosis software
- 21 CFR 892.2100 (DEN190040): Radiological acquisition and/or optimization guidance system

The special controls associated with these regulations are intended to mitigate the risks to health associated with these types of devices. As such, many of the special controls included in these classifications relate directly to elements associated with machine learning-based software devices intended for use in diagnostics. For example, several of the regulations include special controls related to the description of the image analysis algorithm (e.g., 21 CFR 892.2060(b)(1)(i), 21 CFR 870.2785(1), 21 CFR 876.1520(5)). Many others specify elements of the performance testing and characterization. Often included in these regulations (e.g., 21 CFR 892.2060, 21 CFR 892.2070) are special controls that indicate performance must demonstrate that the device provides improved performance on a particular diagnostic task (e.g., detection, diagnosis). For new devices, these requirements generally mean FDA will require both standalone testing characterizing device performance and clinical testing demonstrating diagnostic improvement in the intended use population. For devices implementing machine learning algorithms to estimate other physiologic characteristics, standalone and clinical testing may also be required (e.g., 21 CFR 870.2785). In addition, these regulations may include special controls related to describing the expected performance of the device. Requirements associated with communicating expected device performance in labeling help to (a) mitigate the risks associated with the device and (b) communicate expectations for performance for similar devices to future device developers.

CDRH is statutorily mandated to consider the least burdensome approach to regulatory requirements or decisions. Alternative methods, data sources, real-world evidence, nonclinical data, and other means to meet regulatory requirements may be considered and accepted, when appropriate. FDA encourages innovative approaches to device design as well as mechanisms to address regulatory requirements, when appropriate. FDA takes a benefit–

risk approach to novel devices [5] and to devices with different technological characteristics [6].

CDRH provides opportunities for developers to request feedback and meet with FDA staff to obtain FDA feedback prior to an intended premarket submission [7]. These interactions tend to focus on a particular device and questions relevant to a planned future regulatory submission and may include questions about testing protocols, proposed labeling, regulatory pathways, and design and performance of clinical studies and acceptance criteria.

Device developers need to be aware of all regulatory requirements throughout a product's life cycle including investigational device requirements (e.g., 21 CFR 812), premarket requirements, postmarket requirements (e.g., 21 CFR 820), and surveillance requirements. While this chapter focuses on the premarket and performance assessment of devices, we remind the reader that regulatory requirements throughout the device life cycle should be considered.

3 Study Design and Data Collection

This section aims at summarizing general considerations for study design and data collection for the assessment of ML algorithms in imaging. The specific topics we focus on in this section include study objectives, pilot and pivotal studies, and issues related to data collection, including dataset mismatch and bias. Other study design considerations, such as selection of a reference standard, selection of a performance metric, and data analysis plans, are discussed in later sections.

3.1 Study Objectives

The first consideration in study design is the objective of the study. A general principle is that the study design should aim at generating data to support what the ML algorithm claims to accomplish. The required data are closely related to the intended use of the device, including the target patient population. Important considerations include the significance of information provided by an ML algorithm to a healthcare decision, the state of the healthcare situation or condition that the algorithm addresses, and how the ML algorithm is intended to be integrated into the current standard of care. Examples of study objectives for ML algorithms include standalone performance characterization, standalone performance comparison with another algorithm or device, performance characterization of human users when equipped with the algorithm, and performance comparison of human users with and without the algorithm.

3.2 Pilot and Pivotal Studies

For the purpose of this chapter, a pivotal study is defined as a definitive study in which evidence is gathered to support the safety and effectiveness evaluation of a medical device for its intended use. A pivotal study is the key formal performance assessment of ML devices in medical imaging, and the design of a pivotal study is often the culmination of a significant amount of previous work. An often overlooked, important step toward the design of a pivotal study is a pilot (or exploratory) study. Pilot studies may include different phases, including those that demonstrate the engineering proof of concept, those that lead to a better understanding of the mechanisms involved, those that may lead to iterative improvements in performance, and those that yield essential information for designing a pivotal study. When a pilot study involves patients, sample size is typically small, and data are often conveniently acquired rather than representative of an intended population [8]. Such pilot studies provide information about the estimates of the effect size and variance components that are critical for estimating the sample size for a pivotal study. In addition, a pilot study can uncover basic issues in data collection, including issues about missing or incomplete data and poor imaging protocols. For pivotal studies that include clinicians (typically radiologists or pathologists who interpret images when equipped with the ML algorithm), a pilot study can reveal poor reading protocols and poor reader training [8]. Running one or more pilot studies is therefore highly advisable prior to the design of a pivotal study.

3.3 Data Collection

An important prerequisite for a study that supports the claims of an ML algorithm is that the data collection process should allow the replication of the conclusions drawn from this particular study by independent studies in the future. In this regard, the composition and independence of training and test datasets and dataset representativeness are central issues.

3.3.1 Training and Test Datasets

Training data are defined as the set of patient-related attributes (raw data, images, and other associated information) used for inferring a function between these attributes and the desired output for the ML algorithm. During training, investigators may explore different algorithm architectures for this function and fine-tune the parameters of a selected architecture. The algorithm designer can also partition this data into different sets for preliminary (or exploratory) performance analysis, utilizing, for example, cross-validation techniques [9]. Typically, these cross-validation results are used for further model development, model selection, and hyperparameter tuning. In other words, cross-validation is typically used as an informative step before the ML algorithm is finalized. In many machine learning texts, a subset of data left out

for certain parts of algorithm design (e.g., tuning hyperparameters) is referred to as a validation set. In this chapter, we avoid calling this dataset as a validation set and call it a tuning dataset because it contradicts with the commonly used meaning of validation as "checking the accuracy of" and the definition of validation in 21 CFR 820.3[7] as "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled." Since cross-validation estimates described above are typically used to modify the trained algorithm, they do not pertain to the finalized production version of the ML algorithm.

Test data are defined as the set of patient-related attributes that are used for characterizing the performance of an ML algorithm and performing appropriate statistical tests. For imaging ML software, the performance is estimated by comparing either the output of the finalized software or the interpretation of a human observer who utilizes the software to a reference standard for each case and summarizing the results for the entire dataset using appropriate metrics.

Collecting a well-characterized and representative dataset is resource-intensive, and therefore, most datasets in medical imaging are much more limited in size, compared to, for example, datasets in natural imaging or electronic health records. A general principle for dataset size is that the training dataset should be large enough to minimize overfitting and the test dataset should be large enough to provide adequate precision in testing, including adequate study power when hypothesis testing is involved. Multiple studies have shown that as the training set is gradually increased starting from a small size, overfitting is initially decreased dramatically, with diminishing returns as the dataset size gets larger [10, 11]. The size for which adding more data provides only diminishing returns depends on the complexity of the ML system 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 is extremely important for this task.

3.3.2 *Independence*

A central principle in performance assessment is that the test dataset is required to be independent of the training dataset, meaning that the data for the cases in the test set do not depend on the data for the cases in the training set. It is well-known that the violation of this principle results in optimistically biased performance estimates [12]. To avoid this bias, developers typically set aside a dedicated test data for performance estimation aimed to be independent of the training dataset. There are subtle ways in which the independence principle can be violated if the test dataset is not carefully

[7] https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=820.3

selected. We discuss two such mechanisms below. The first is related to including data from a particular patient in both the training and test datasets. The second is related to performing internal validation instead of using an external validation [13] method.

A basic mechanism that can cause a dependence between the training and test sets is the inclusion of data from one patient in both datasets. This could happen if different regions of interest, different image slices, or different objects from the same patients span both the training and test datasets. Since portions of the data from the same patient are expected to be correlated, this practice will result in a statistical dependence between the training and test datasets. A straightforward principle to be followed is to include each patient's data exclusively in the training set or exclusively in the test set.

A more subtle mechanism that can cause a dependence between the training and test datasets is the way the data are sampled or the way that one dataset is partitioned into training and test datasets. Internal validation, which involves partitioning a previously collected sample into training and test datasets randomly or in a stratified way across a given attribute, may result in a dependence between the training and test datasets. Any sampled data, even if it was designed to be collected in a random manner, may not perfectly follow the true distribution of the target population due to finite sample size effects. In addition, there may be a systematic deviation in the feature distribution of a particular sample from the true distribution due to the fact that, for example, the sample may be collected only at a particular site or using only a particular or predominant image acquisition system that does not represent the true distribution. When such a dataset is shuffled and randomly partitioned into training and test datasets, knowledge about the distribution of the training data may provide unfair information about the distribution of the test dataset that would have been impossible to know had the training and test datasets been sampled independently from the true population. A practical approach to reduce this type of dependence is to sample the training and test datasets from multiple different, independent sites, a practice known as external validation [13].

3.3.3 Representativeness

ML algorithms are data-driven, and the distributions of the training and test data have direct implications for algorithm performance and its measurement. Ideally, training and test sets should be large and representative enough so that the collected data provides a good approximation to the true distribution in the target population. As discussed above, well-characterized and annotated medical imaging datasets are typically limited in size. When the dataset size is a constraint, informativeness of a case to be selected for training

the ML algorithm for the task at hand is an additional consideration besides representativeness [14]. Active machine learning techniques aim to proactively select training cases that can best improve model performance, based on informativeness, representativeness, or a combination of the two [15]. Active learning techniques have been applied to train ML algorithms applied to brain imaging [16, 17].

Representativeness of the test dataset is typically desirable when an unbiased estimate of the ML algorithm performance assessment is sought for the target population. For most classification problems, representativeness within each class may be sufficient, which allows designers to enrich the test datasets with classes that have smaller prevalence in the target population. For studies that aim to compare two competing arms (e.g., clinicians' image interpretation with and without ML), enrichment methods that are based on a measurement (e.g., patient or lesion characteristics, risk factors), which trade the unbiased absolute performance results for the practical ability to compare the two competing arms with possible moderate biases, are often acceptable [8]. For example, if cases that are known to be trivial to classify (or diagnose) in both arms of a comparative study are excluded from the test dataset, this will result in a bias in the absolute performance estimates for both arms but may not result in a bias in the difference or change the ranking order of the two arms under comparison, thus allowing the use of a smaller test dataset and a less resource-intensive study design. Likewise, as discussed in Subheading 6.5, when the main goal is to compare the standalone performance of two algorithms to determine which algorithm or modification performs best, it is possible to perform the comparison on a smaller enriched dataset with a careful sampling strategy that does not result in a bias in the difference of the two performance estimates.

3.3.4 *Dataset Mismatch*

Dataset mismatch is defined as a condition where training and test data follow different distributions, which is popularly known as "dataset shift" in the ML literature [18]. We prefer using "mismatch" because "shift" specifically refers to adding a constant value to each member of a dataset in probability distribution theory, which does not convey all types of mismatches that the term is intended for. Dataset mismatch can also be between test data and real-world deployment data (rather than test and training) or current real-world data vs. future real-world data (e.g., due to changes in clinical practice). There may be many potential reasons for dataset mismatch, with sample selection bias and non-stationary environments cited as the most important ones [19]. Storkey [20] grouped these mismatches into six main categories, including sample selection bias, imbalanced data, simple covariate shift, prior probability shift, domain shift, and source component shift. Dataset

mismatch may result in poor performance of the trained ML algorithm. In addition, especially if caused by a non-stationary environment, dataset mismatch may mean that the performance assessment results obtained at premarket testing may no longer be valid in the clinical environment. A first step in mitigating the effects of dataset mismatch is to detect it. Several methods, including those based on distance measures [21] and dimensionality reduction followed by statistical hypothesis testing [22], have been proposed for this purpose. Techniques for mitigating the effect of dataset mismatch include importance weighting [23] and utilizing stratification, cost curves, or mixture models [24], among others.

3.4 Bias

Bias is a critical factor to consider in study design and analysis for ML assessment, and here we intend to give an overview of sources of bias in ML development and assessment. Note that the general artificial intelligence and machine learning literature currently lacks a consensus on the terminology regarding bias. We consider that performance assessment of an ML system from a finite sample can be cast as a statistical estimation problem. In statistics, a biased estimator is one that provides estimates which are systematically too high or too low [25]. Paralleling this definition, we define statistical bias as a systematic difference between the average performance estimate of an ML system tested in a specified manner and its true performance on the intended population. This systematic difference may result from flaws in any of the components of the assessment framework shown in Fig. 1: collection of patient data and the definition of a reference standard (for both algorithm design and testing stages), algorithm training, analysis methods, and algorithm deployment in the clinic.

Note that the definition of statistical bias above includes systematically different results for different subgroups. ISO/IEC Draft International Standard 22,989 (artificial intelligence concepts and terminology) defines bias as systematic difference in treatment of certain objects, people, or groups in comparison to others, where treatment is any kind of action, including perception, observation, representation, prediction, or decision. As such, statistical bias may result in the type of bias defined in the ISO/IEC Draft International Standard.

We start our discussion of bias with the effect of the dataset representativeness, which has direct implications for ML algorithm performance and its measurement, as described above. When the dataset is not representative of the target population, this can lead to *selection bias*. For example, if all the images in the training or test datasets are acquired with a particular type of scanner while the target patient population may be scanned by many types of scanners, this may lead to an ML algorithm performance estimate that is systematically different from that on the intended population or lead to different results for different subgroups. *Spectrum bias*,

which can be viewed as a consequence of selection bias, describes a systematic error in performance assessment that occurs when the sample of cases studied does not include a complete spectrum of patient and disease characteristics [26]. *Imperfect reference standard bias* and *verification bias* are two types of biases that are related to the reference standard (Subheading 5); the former applies to conditions in which the reference standard procedure is not 100% accurate, and the latter applies to conditions in which only subjects verified for presence or absence of the condition of interest by the reference standard are included in the training set or test set.

Aggregation bias and model design bias are two types of biases that can occur in the algorithm training stage. *Aggregation bias* is related to the information loss which occurs in the substitution of aggregate, or macro-level, data for micro-level data. Aggregation bias can lead to a model that is not optimal for any group or a model that is fit to the dominant population [27]. In ML architecture selection and algorithm training, the designer often has options for model design that may affect the objectives of accuracy, robustness, and fairness, and these objectives may have intrinsic trade-offs. *Model design bias* refers to the design choices that may amplify performance disparities among minority and majority data subgroups [28].

In addition to biases stemming from test dataset composition and the reference standard discussed above, inappropriate selection of the performance metric in the data analysis stage may result in a bias. Many metrics used for evaluation of image analysis algorithms, such as the mean squared error (MSE) for image noise reduction, do not represent the task that the ML algorithm was designed for, e.g., the detection of low-contrast objects in a noisy image. The use of an inappropriate metric may thus result in a difference between the test and true performance, e.g., a conclusion that the algorithm is helpful for its intended use when in clinical reality it is not.

Several factors may contribute to bias after a medical ML system is introduced into the clinic. One of these is the *bias due to a temporal dataset shift* [29] that may cause a mismatch between the data distribution on which the system was developed/tested and the distribution to which the system is applied. Another type of bias, sometimes termed *deployment bias* [27], may be caused by the use of a device in a manner that was not tested as part of the performance assessment and hence does not conform with the intended use of the device, e.g., off-label use. Other types of biases during deployment are also possible because of the differences in the test and clinical environments and unanticipated issues in the integration of the ML system into clinical practice.

4 Algorithm Documentation

4.1 Why Algorithm Documentation Is Important

Machine learning (ML) algorithms have been evolving from traditional techniques with hand-crafted features and interpretable statistical learning models to the more recent deep learning-based neural network models with drastically increased complexity. Appropriate documentation of ML algorithms is critically important for reproducibility and transparency from a scientific point of view. Algorithm description with sufficient details is particularly important in a regulatory setting reviewed by regulators for the assessment of technical quality, for comparing with a legally marketed device, and for the assessment of changes of the algorithm in future versions.

Reproducibility is a well-known cornerstone of science; for scientific findings to be valid and reliable, it is fundamentally important that the experimental procedure is reproducible, whether the experiments are conducted physically or in silico. ML studies for detection, diagnosis, or other means of characterization of brain disorders or other diseases are in silico experiments involving complex algorithms and big data. As such, we adopt the definition of reproducibility from a National Academies of Sciences, Engineering, and Medicine report [30] as "obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis." It has been widely recognized that poor documentation such as incomplete data annotation or specification of data processing and analysis is a primary culprit for poor reproducibility in many biomedical studies [31]. Lack of reproducibility may result in not only inconvenience or inferior quality but sometimes a flawed model that can bring real danger to patients when such models are used to tailor treatments in drug clinical trials, as reported by Baggerly and Coombes [32] in their forensic bioinformatics study on a model of gene expression signatures to predict patient response to multidrug regimens.

Appropriate documentation of algorithm design and development is essential for the assessment of technical quality. Identification of the various sources of bias discussed in Subheading 3.4 may not be possible without appropriate algorithm documentation. Furthermore, while there is currently no principled guidance on the design of deep neural network architectures, consensus on good practices and empirical evidence do provide basis for the assessment of technical soundness of an ML algorithm. For example, the choice of loss function is closely related to the clinical task: mean squared error is appropriate for quantification tasks, cross-entropy is often used for classification tasks, and so on. Moreover, the design and optimization of algorithms involve trial-and-error and ad hoc procedures to tune parameters; as such, a developer may introduce bias even unconsciously if the use of patient data and truth labels is not properly documented.

Documentation of ML algorithms is often necessary in a regulatory setting and generally required by FDA under 21 CFR 820.30 design controls. As mentioned in Subheading 2.1, comparison of technological characteristics of a premarket device with a legally marketed predicate device is one of the essential components in determining substantial equivalence for a 510(k) submission. Moreover, the ML algorithm in an FDA-authorized device is often updated, and appropriate algorithm documentation is crucial to decide if a new version has undergone major updates that would require a re-submission to the FDA.

4.2 Essential Elements in Algorithm Description

Many efforts in academia have been devoted to developing checklists for ML algorithm development and reporting to enhance transparency, improve quality, and facilitate reproducibility. A report from the NeurIPS 2019 Reproducibility Program [33] provided a checklist for general machine learning research. Norgeot et al. [34] presented the minimum information about clinical artificial intelligence modeling (MI-CLAIM) checklist as a tool to improve transparency reporting of AI algorithms in medicine. The journal *Radiology* published an editorial with a checklist for artificial intelligence in medical imaging (CLAIM) [35] as a guide for authors and reviewers. Similarly, an editorial of the journal *Medical Physics* introduced a required checklist to ensure rigorous and reproducible research of AI/ML in the field of medical physics [36]. Consensus groups also published the SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) as guidelines for clinical trial protocols for interventions involving artificial intelligence [37]. Also, there are undergoing efforts on guidelines for diagnostic and predictive AI models such as the TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Machine Learning) [38] and STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence).

Besides the abovementioned references, the FDA has published a guidance document for premarket notification [510(k)] submissions on computer-assisted detection devices applied to radiology images and radiology device data [39]. Here we provide a list of key elements in describing ML algorithms for medical imaging applications, which we believe are essential (but not necessarily complete) for understanding and technical assessment of an ML algorithm.

- Input.
  The types of data the algorithm takes as input may include images and possibly non-imaging data. For input data, essential information includes modality (e.g., CT, MRI, clinical data), compatible acquisition systems (e.g., image scanner manufacturer and model), acquisition parameter ranges (e.g., kVp range, slice thickness in CT imaging), and clinical data collection protocols (e.g., use of contrast agent, MRI sequence).

- Preprocessing.
  Input images are often preprocessed such that they are in a suitable form or orientation for further processing. Preprocessing often includes data normalization, which refers to calibration or transformation of image data to that of a reference image (e.g., warping to the reference frame) or to certain numerical range, e.g., slice thickness normalization. Other examples of preprocessing include elimination of irrelevant structures such as a head holder, image size normalization, image orientation normalization, and so on. Sometimes an image quality checker is applied to exclude data with severe artifacts or insufficient quality from further processing and analyses. It is important to describe the specific techniques for normalization and image quality checking. Furthermore, it is critical to make clear how cases failing the quality check are handled clinically (e.g., re-imaging or reviewed by a physician) and account for the excluded cases in the performance assessment.

- Algorithm architecture.
  Algorithm architecture is the core module of a machine learning algorithm. In traditional ML techniques, hand-crafted features that are often motivated by physician's experiences are first derived from medical images. A feature selection procedure can be applied to the initially extracted features to select the most useful features for the clinical task of interest. The selected features are then merged by a classifier into a decision variable. There are many choices of the classifier depending on the nature of the data and the purpose of the classifier: linear or quadratic discriminant analysis, k nearest neighbor (kNN) classifiers, artificial neural networks (ANNs), support vector machines, random forests, etc. As such, the algorithm description typically includes the definition of features, the feature selection methods, and the specific classification model. Moreover, it is important to document hyperparameters and the method with which these hyperparameters are determined, for example, the number of neighbors in the kNN method, the number of layers in ANNs, etc.

Recently, deep learning neural network algorithms have been widely used in medical imaging applications. The NN architecture in this type of ML algorithms is composed of a large number of layers that learn to represent data at multiple levels of abstraction and automatically learns features from raw medical image data. As such, instead of sequential hand-crafted feature extraction, selection, and classification, automatic feature engineering and classification (or other types of decision-making such as quantification) are seamlessly integrated in one deep NN architecture. If a published architecture such as AlexNet, VGGNet, Inception V3, etc. is followed exactly, a succinct description is to refer to the reference. Otherwise, the architecture is typically described using a diagram with details such as the number and type of layers, the number of nodes in each layer, the activation functions, the loss function, and so on.

Sometimes hand-crafted features are combined with CNN-based automatic features by a traditional classifier (e.g., random forest) to take advantage of both the power of deep learning in information extraction and domain-specific expertise. In this situation, architecture description includes the entire pipeline, both types of information as described in the above two paragraphs.

- Algorithm Training.
  ML algorithm training is the process of designing ML algorithm architecture, optimizing the parameters, and selecting the hyperparameters. Taking the popular deep neural networks as an example, the first step in training is to design an architecture or adopt one that has been proven successful in similar applications (see previous bullet). Parameters mainly refer to network weight and bias parameters for combining node outputs in one layer as inputs to nodes of the next layer. Hyperparameters include both those related to network architecture and those related to parameter optimization strategies. Network architecture hyperparameters such as number of hidden layers and units can be pre-selected and fixed if an established architecture is adopted and/or further tuned during training. Another architecture hyperparameter that has been popularly used to avoid overtraining is dropout rate, which refers to the probability of a neuron being "dropped out" in a training step (i.e., the weights are not updated) but may be active in the next step. Hyperparameters related to parameter optimization include learning rate, momentum, number of epochs, batch size, etc.

  Given a set of hyperparameters, the network parameters are optimized using training images and associated truth labels. The hyperparameters are typically tuned using a separate tuning dataset. *See* Subheading 3 for discussion of training data. Again, it is important to fully describe the training process and training data as part of the algorithm description.

- Post-processing and Output.
  Given the output of the main ML algorithm, some post-processing steps may be followed, for example, to transform the output to a more interpretable form. The final outputs of an ML algorithm are those that are presented to the end user such as a radiologist or other clinicians. They can be marks on the images indicating the algorithm-determined suspicious areas, a quantitative score indicating the algorithm-estimated likelihood of disease severity, and/or a binary classification indicating if the lesion is benign or malignant, etc. The algorithm description must make clear the final algorithm outputs and how they are intended to be used clinically so that appropriate validation and testing studies can be conducted.

Finally, it should be emphasized that a great description of ML algorithms not only provides these essential elements but also, more usefully, provides rationale on the algorithmic choices. Such rationale may include established good machine learning practices, evidence from similar applications, or methodological research that helps avoid overfitting, reduce bias, and improve generalizability.

5 Reference Standard

Rigorously developed, well-accepted reference standards (also called the "gold standard" or "ground truth") for training and evaluating machine learning algorithms are essential to validating and characterizing the performance of machine learning algorithms. The reference standard provides a definitive or quasi-definitive characterization of the case based on information that may not be part of the machine learning input, such as biopsy or 1-year follow-up for radiological imaging oncology applications (for an example in the regulatory setting, *see*[8]). The "truthing" procedures for the cases included in validation (especially external validation) should utilize the best reference standard as recognized by the scientific community to help ensure that the performance of the device is well-characterized. The truthing process is distinct from other aspects of evaluating ML performance as the goal is to determine the "correct" characterization of each case, not to evaluate the device and reader performance in assessing a particular case.

Brain disorders often represent unique challenges to establishing appropriate reference standards. Generally, reference standard can be based on established clinical determination (including an independent modality recognized as a gold standard), follow-up clinical examination, or follow-up medical examination other than imaging. For brain disorders, the pathophysiology may be poorly

[8] https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf

understood, the progression of some disease may be slow making it difficult to reliably observe changes over time, the clinical definition of the condition may rely heavily on subjective assessments, or definitive assessments may be delayed by years (e.g., Alzheimer's disease) with different syndromes mistaken for the condition of interest (e.g., Parkinsonian-like disorders). In other words, for some brain conditions, the current best available reference standard is based on clinical determination, and confirmation from an alternative method (for instance, histopathological confirmation) may be desirable. Furthermore, ethical and pragmatic challenges of obtaining neurological tissue samples that would allow for independent pathological assessment may limit the utility of biopsy or tissue resection in many brain disorders.

In limited instances, alternatives to independent confirmation of the case "truth," such as interpretation by a reviewing clinician (s), may be considered. Especially in brain disorders where the diagnostic criteria may already be challenging, the importance of multiple reviewing clinicians using the best possible information, even if that requires long-term follow-up, cannot be understated. For some brain disorders such as chronic traumatic encephalopathy or Alzheimer's disease, outstanding challenges remain as the reference standard may be best assessed by biopsy or following death (i.e., autopsy). Greater biological and physiological understanding may be needed to inform the correct diagnosis early in disorder development. Using machine learning techniques to assist in this process is tempting, but the performance will generally be limited by the correctness of the reference standard. In other words, how would we assess if the ML device is outperforming the reference standard as any disagreement may be considered incorrect based on the reference truth?

Uncertainty in the reference standard needs to be accounted for in the analysis. For some machine learning devices, reference standard by expert assessment can be considered, depending on the indications for use, intended use, benefit–risk profile, and device outputs. This is often the case of ML algorithms used in segmentation tasks. In these limited instances, the reference standard from a single clinical truther remains undesirable due to potential concerns about bias or the overall performance of the truther (that is, they are not likely to be 100% accurate, especially for challenging cases). Therefore, multiple clinical truthers are desired. Truthing processes using top experts or truthing processes that weight the clinical truthers' "accuracy" in the construction of the reference standards may also be considered (e.g., *see* Warfield et al. [40]).

When the truthing process involves interpretation by a reviewing clinician, the number of truthers; their qualifications, experience, and expertise; the instructions for the truthing process; and any other information should be described and documented. In instances where multiple truthers are involved, developer must

consider in advance how the interpretations of these various readers will be incorporated into the final study design and analysis. While combining the interpretation of all truthers into a single reference truth for a particular case may be appropriate in some instances, in other cases such as when the variability between truthers is high, study designs and data analysis methods that take into consideration variability in the reference standard may be appropriate. For instance, reference standard by panel discussions may face unique challenges, especially when loud voices, biases, and group dynamics may influence the outcome. On the other hand, majority vote may lead to other biases such as in segmentation where only including voxels from the majority could lead to small areas or volumes even when compared to all of the participating truthers.

Certain practices in development of the reference standard should be avoided. Often developers look for reference standards of convenience such as a single truther observing the same input data, such as a CT image, as the machine learning algorithm inputs and providing their best judgment as the underlying "truth" of the case. Truthers should not be used as readers who read those images as part of the evaluation of device performance because that can introduce considerable bias to the study results. Public data with unclear processes for establishing the reference standard or incomplete case-level data (such as data without follow-up information, without other typically assessed test results, or incomplete demographic information) frequently raises concerns about the appropriateness of the reference standard in these instances.

The reference standard should generally be based on the best available evidence for the case as recognized by the scientific community. The goal of the reference standard is to establish the "truth" for the outcome of the case. This may present challenges to cohorts where the amount of evidence may differ between cases. Requirements for the minimal amount of information available for a particular case to establish a reasonable "truth" should be defined in advance in the premarket and postmarket setting. As with overall device classification, expectations for rigor and certainty in the reference standard may increase with the device risk associated with misclassification or misdiagnosis. In a regulatory context, often more flexibility is generally permitted in the reference standard for the training data as compared to expectations of rigor in the reference standard for the validation data. Finally, the use of synthetic data is attractive as these techniques provide some opportunities for more well-characterized reference standards in some applications. While synthetic data presents an intriguing approach to addressing some challenges related to reference standards in brain disorders, this is a fairly new topic without significant experience within the current regulatory framework.

6 Standalone Performance Assessment

ML standalone performance is a measure of algorithm performance independent of any human interaction with the ML tool [41]. Standalone performance is the primary assessment for autonomous ML tools that make decisions without clinicians' interactions but may be only one element of assessment for an ML algorithm used as an aid, in which case a clinical assessment of reader performance utilizing the ML may also be required, as discussed in Subheading 7. Standalone testing is also used heavily during algorithm development to benchmark performance and compare potential algorithm modifications before a "final" version is determined. This is because it is often straightforward to integrate iterative testing within the development framework. Standalone testing spans a wide range of possible implementations from initial validation of modifications using a small dataset through large-scale evaluations across multiple independent sites [42] which provides a higher level of confidence in algorithm performance.

Sometimes researchers assume that standalone testing is not important, or at least not as important, as a clinical evaluation, especially for ML-assist devices. However, standalone testing is critical even when a clinical reader study is performed because it is often conducted on larger and more diverse datasets allowing for more refined subgroup analyses and understanding of performance characteristics. It is also critical for assessing the robustness of an ML algorithm and for comparing performance across different algorithms.

In the following, we describe study design, study endpoints, and approaches for assessing standalone performance for specific types of ML tools.

6.1 Segmentation Assessment

Accurate segmentation of brain structures is routinely used in many neurological diseases and conditions when imaging with modalities such as CT, MRI, and PET. As an example, quantitative analysis of brain MRI has been used in assessing brain disorders such as Alzheimer's disease, epilepsy, schizophrenia, multiple sclerosis (MS), cancer, and infectious and degenerative diseases [43]. Often brain assessment quantifying change over time requires the segmentation of brain tissue or anatomy. We define segmentation as the process of partitioning a brain image or image volume into multiple objects defined by a set of voxels unique to each structure or object of interest.

There have been various methods proposed for assessing how well an ML algorithm characterizes objects and how one segmentation algorithm compares to another. Zhang discusses three basic approaches to assessing segmentation algorithms in general [44]. This includes analytical methods, goodness methods, and

discrepancy methods [45]. Analytical methods consider the principles, requirements, utilities, and complexity of segmentation algorithms but can be quite difficult to apply, especially to DL-based segmentation because not all algorithm properties are easily obtained. Goodness methods evaluate segmentation performance by judging the segmented images based on certain quality measures established according to human intuition and include measures such as inter-region uniformity, inter-region contrast, and region shape [45]. Discrepancy methods quantify the difference between segmented objects and a reference standard segmentation. They are the most common type for assessing segmentation algorithms with the caveat that often a ground-truth segmentation is not available. In this case, algorithm segmentations are then compared to human segmentations where the human segmentation is considered the reference standard. Since human segmentations of brain anatomy and structure can be quite variable, segmentations by multiple truthing readers are often collected, and an aggregated reference [40, 46] is used, or the agreement or interchangeability of the algorithm with a truthing reader is assessed.

The remainder of this subsection describes a few common segmentation metrics where we assume a hard segmentation and a single reference standard. A hard segmentation means a voxel is either part of the segmentation or not (this is in contrast to a soft or fuzzy segmentation which means that each voxel is assigned a probability of being part of the segmentation).

An example of a 2D segmentation $\boldsymbol{S}$ and reference segmentation $\boldsymbol{R}$ is shown in Fig. 2 for image $\boldsymbol{X}$. The false-positive (FP), true-positive (TP), false-negative (FN), and true-negative (TN) regions are also shown. Taha and Hanbury provide a nice overview of 20 segmentation metrics used for discrepancy assessment [47]. Please refer to this paper for more details on many segmentation assessment approaches including methods for assessing fuzzy segmentation algorithms [47]. We next discuss some of the discrepancy assessment approaches frequently used in the literature.

Overlap indexes assess a segmentation by how well it overlaps with the reference. We define some basic overlap metrics below using TP, TN, FP, and FN as voxel counts in the definitions.

- Voxel true-positive rate (TPR), sensitivity, recall: proportion of correctly segmented reference voxels.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- Voxel true-negative rate (TNR), specificity: proportion of correctly segmented background voxels.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

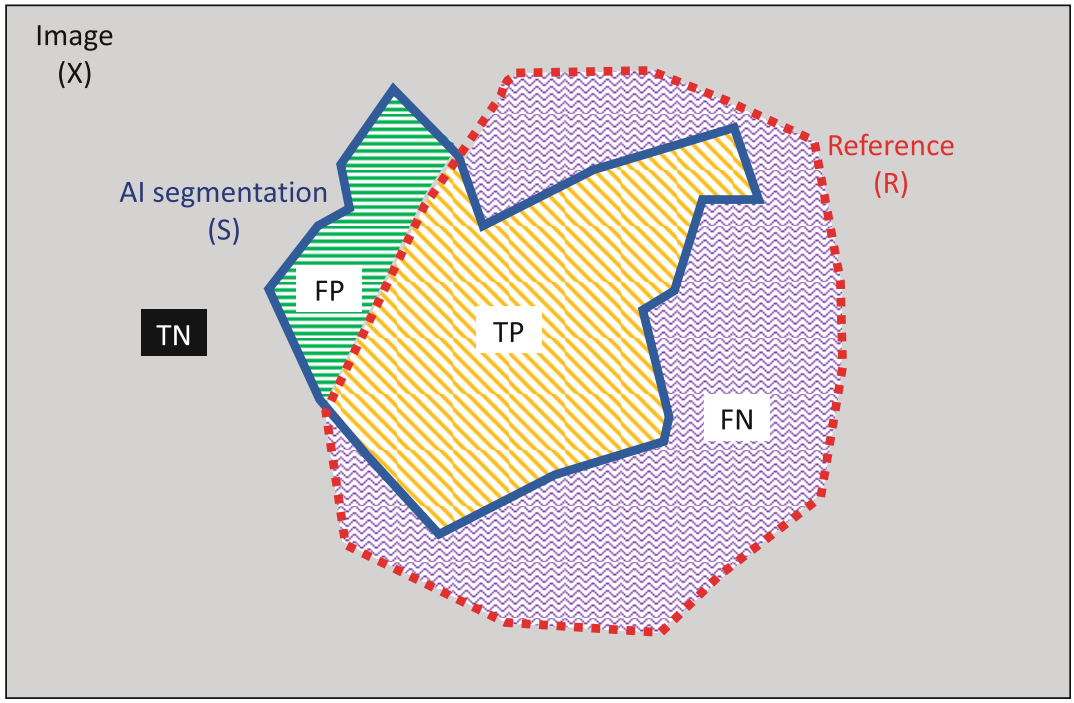

Fig. 2 Diagram of a segmented object (blue solid line) overlaid by the reference segmentation (red dashed line). The false-positive (FP) voxels (green hashed region), true-positive (TP) voxels (yellow hashed region), false-negative (FN) voxels (purple wave region), and true-negative (TN) voxels (gray region) are shown in the figure as well

- Voxel accuracy [45]: proportion of correctly segmented voxels (including both reference and background voxels).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Dice similarity coefficient (DSC), F1 metric [48]

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = \frac{2\text{JI}}{1 + \text{JI}}$$

- Jaccard index (JI), intersection over union (IoU) metric [49]

$$\text{JI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} = \frac{\text{DSC}}{2 - \text{DSC}}$$

The Dice coefficient (DSC) is the most widely used performance metric for characterizing medical image volume segmentations including brain segmentations and can also be used to assess the reproducibility of multiple annotations [47]. The Jaccard index is another common assessment metric. JI and DSC are monotonically related with DSC always having a larger value than JI except at 0 and 1 when the two are equal. However, they have different

properties when averaging performance across multiple segmentations where Jaccard penalizes large segmentation errors more than Dice (somewhat similar to how an L2 norm penalizes larger error more than an L1 norm) [50].

Accuracy is another commonly reported metric, but accuracy is often dominated by a large disparity in the number of reference and background voxels within an image. Accuracy can be high even for poor overlap in the segmentation and reference when the vast majority of voxels in the image are background. This can make it difficult to differentiate between algorithms based only on accuracy differences. A similar observation can be made for specificity or, more generally, for any segmentation metrics involving TN. Indeed, since most voxels are background, TN can be very large. Finally, the definition of the background is not always straightforward and can sometimes be arbitrary (for instance, if the background depends on the field of view of the image).

Distance-based metrics are useful when the boundary of the segmentation is critical [51]. They assess the distance between the segmentation boundary and the reference boundary taking into account the spatial position of the boundary voxels [47]. Some common distance metrics include:

- *Hausdorff distance (HD)* between two voxel sets B_S and B_R (sets of boundary voxels) [47]

  $HD = \max\left(h(B_S, B_R),\ h(B_R, B_S)\right)$, where $h(B_R, B_S) = \max_{r \in B_R} \min_{s \in B_S} \|r - s\|$
- *Mahalanobis distance (MHD)* between two voxel sets B_S and B_R.

  $MHD = \sqrt{\left(\mu_{B_S} - \mu_{B_R}\right) S^{-1} \left(\mu_{B_S} - \mu_{B_R}\right)}$, where μ_{B_S} and μ_{B_R} are the means of the point sets and S is the common covariance matrix of the two sets [47]

There are additional segmentation assessment metrics including volume metrics, information theoretic metrics (e.g., mutual information), probabilistic metrics (e.g., intraclass correlation coefficient [ICC]), and pair counting metrics that can also be used to assess the quality of a segmentation algorithm or for comparing multiple segmentations [47].

6.2 Classification Assessment

Classification ML are algorithms designed to parse brain images and data into unique categories. Often the task is differentiating two groups (e.g., cancer versus non-cancer patients), but classification can also be multiclass (e.g., differentiating astrocytoma, glioblastoma, and meningioma brain tumors). The outputs of an ML algorithm can be discrete classes (e.g., via decision tree) or a continuous or a quasi-continuous score (e.g., output of a linear classifier and many DL methods) for an image. As with all ML, the classifier output needs to be assessed and properly interpreted, so

ML performance is understood in the correct context. Tharwat [52] and Hossin and Sulaiman [53] have nice summaries of classification analysis methods. They discuss various performance metrics along with information on how and when each metric might be most effectively used in classifier assessment.

In the remainder of this subsection, we concentrate on binary classifier assessment that includes a wide range of statistical metrics for assessing classifier performance starting with operating point metrics defined directly from discrete ML outputs and moving to more complex metrics based on thresholding a continuous ML output score.

Some basic prevalence-independent metrics (i.e., metrics that do not depend on the prevalence of diseased cases in the standalone database) are described below where TP, TN, FP, and FN are case counts here.

- True-positive rate (TPR), sensitivity, recall

$$\mathrm{TPR} = Se = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- True-negative rate (TNR), specificity

$$\mathrm{TNR} = Sp = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Likelihood ratios are aggregate measures combining sensitivity and specificity. The positive/negative likelihood ratio is the ratio of the probability of a person who has the disease testing positive/negative over the probability of a person who does not have the disease testing positive/negative. They are defined as:

- Positive likelihood ratio (LR^+)

$$\mathrm{LR}^+ = \frac{\mathrm{TPR}}{1 - \mathrm{TNR}}$$

- Negative likelihood ratio (LR^-)

$$\mathrm{LR}^- = \frac{1 - \mathrm{TPR}}{\mathrm{TNR}}$$

Other operating point metrics depend on the prevalence of disease in the test dataset. They include:

- Positive prediction value (PPV), precision

$$\mathrm{PPV} = \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}}$$

- Negative prediction value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{FN} + \mathrm{TN}}$$

- Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- F_1 score

$$F_1 = 2 \cdot \frac{\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}}$$

These metrics are most appropriate when assessing ML performance in a dataset representing the true clinical population because of their prevalence dependence. They must be interpreted with caution when applied to enriched datasets especially when extrapolating the estimated classification performance to the clinical environment.

For continuous ML scores where a final classification is based on applying a threshold to the output scores, there are aggregation measures that more completely characterize overall classifier performance for a binary task. A common choice is receiver operating characteristic (ROC) analysis which characterizes performance for all possible operating points of the classifier. An ROC curve plots TPR as a function of the false-positive rate (FPR = 1-TNR) when the threshold on the classifier output is varied over the complete range of possible output scores [54–56]. An example of an ROC curve is shown in Fig. 3. The advantage of the ROC curve is it shows the benefit (i.e., TPR) as a function of all possible risk values (i.e., FPR) such that a much more complete understanding of the benefit–risk trade-off at all operating points is provided [52].

To facilitate statistical comparisons and to benchmark performance, summary performance can be estimated from ROC curves with the most popular being the area under the ROC curve (AUC) and the partial AUC (PAUC area under just a portion of the ROC curve) [41]. However, the ROC curve should always be plotted to allow for a visual assessment of an individual algorithm's performance or to facilitate a comparison across algorithms. This allows the trade-off across the full range of the ROC curve to be visualized.

Parametric and nonparametric statistical methods are available to both estimate AUC/PAUC and their uncertainties. These approaches allow for statistical comparisons in performance among multiple ML algorithms. There is substantial literature on statistical method for assessing and comparing ROC performance. A great summary of approaches can be found in a report on ROC by the International Commission on Radiation Units and Measurements (ICRU) [57].

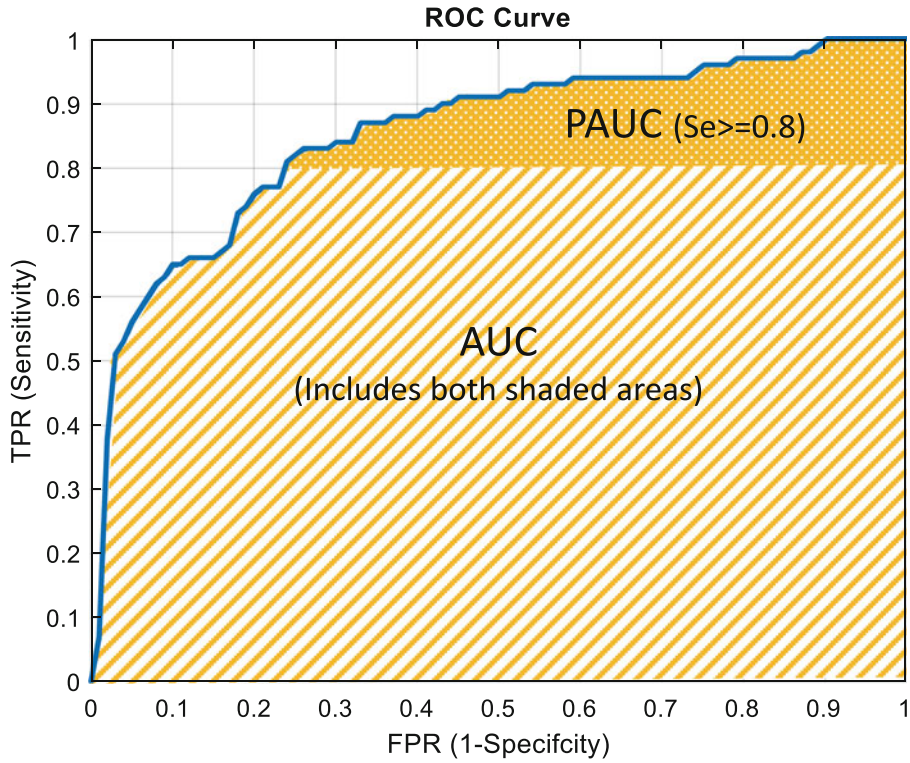

Fig. 3 Plot of an ROC curve for an ML algorithm in a binary classification task. The ROC curve (blue line) shows the trade-off between sensitivity (TPR) and specificity (FPR) for all possible operating points. Both the AUC (includes both shaded regions) and the PAUC for sensitivity $\geq$0.8 are shown in the figure

6.3 Abnormality Detection Assessment

ML detection algorithms mark locations or regions of an image that may reveal abnormalities [41]. Examples of basic ML detection include ML-based bounding boxes or segmentations of potential brain lesions or markers indicating potential brain lesions in an MR or CT scan. Often ML detection outputs include not only localization information but also a confidence score or class determination for the identified regions such that the ML includes both detection and classification functionalities. In the remainder of this subsection, we concentrate on assessing only detection performance without addressing any other potential components of an ML algorithm's output. However, we will still use the ML detection confidence scores, when available, to expand the range of possible performance metrics available for standalone assessment.

Similar to classification metrics, there are a wide range of metrics available for assessing detection performance. Basic detection operating point metrics, usually based on thresholding a continuous ML score for each region, include counts of object-based true-positive (TP), false-positive (FP), and false-negative (FN) detections using the basic definitions in Subheading 6.2 above. Note that object-based true-negative (TN) detections are

generally not estimable in ML detection because there is an infinite (or at least an extremely larger number) of possible TN locations within an image [41]. In addition, ML detection assessment is complicated by the need for a predefined rule (method and threshold) for determining a "correct" detection based on the overlap of a bounding box/segmentation with a reference standard region or the distance from an ML marker to a reference standard object (e.g., distance to the centroid of a reference standard). The overlap metric is often based on the intersection over union (IoU) for bounding boxes/segmentations with a reference standard object and Euclidean distance for markers. However, other potential overlap metrics and criteria may be justifiable for various detection tasks.

Based on the number of TP, FP, and FN detection counts, some basic summary operating point metrics include the true-positive rate (TPR) (i.e., recall) and positive predictive value (PPV) (i.e., precision) that are defined similarly as those in Subheading 6.2 but with the unit of regions/cases instead of voxels and the number of FPs per case (or another appropriate unit of interest) since individual cases often include multiple images and abnormalities of interest [41]. For example, an MR exam of the head may include multiple MRI sequences (e.g., T1, T2) such that it is possible to report ML detection performance on a per-patient, per-view (sequence), or per-abnormality (object) basis. The unit of performance should be clearly defined and justified with per-abnormality (or object) performance typically being reported for most image-based ML detection devices especially when only a single exam is available per patient.

Analogous to classification tasks, aggregation metrics that more completely characterize overall ML detection performance are used when a confidence score is available for each detection. ROC analysis is not generally used for ML detection assessment because, as mentioned previously, TNs are not estimable. Therefore, alternate methods have been developed including the free-response receiver operating characteristic (FROC) analysis. FROC accounts for localization and detection of an arbitrary number of abnormalities within an image set [58]. FROC curves plot the fraction of correctly localized lesions as a function of the average number of FPs across the full range of confidence scores for an ML detection algorithm [59]. An example of FROC curve is shown in Fig. 4.

The plot in Fig. 4 shows a nonparametric FROC curve. Parametric FROC methods have been developed using maximum likelihood methods [60–62]. Similar to ROC analysis, FROC area-based metrics can serve as summary performance metrics, but, since the number of FPs in FROC are not bounded, the area under the curve is not limited. This complicates the use of the full area under the FROC curve as a summary figure of merit. Therefore, alternate area-based figures of merit have been developed to summarize and compare FROC performance curves.

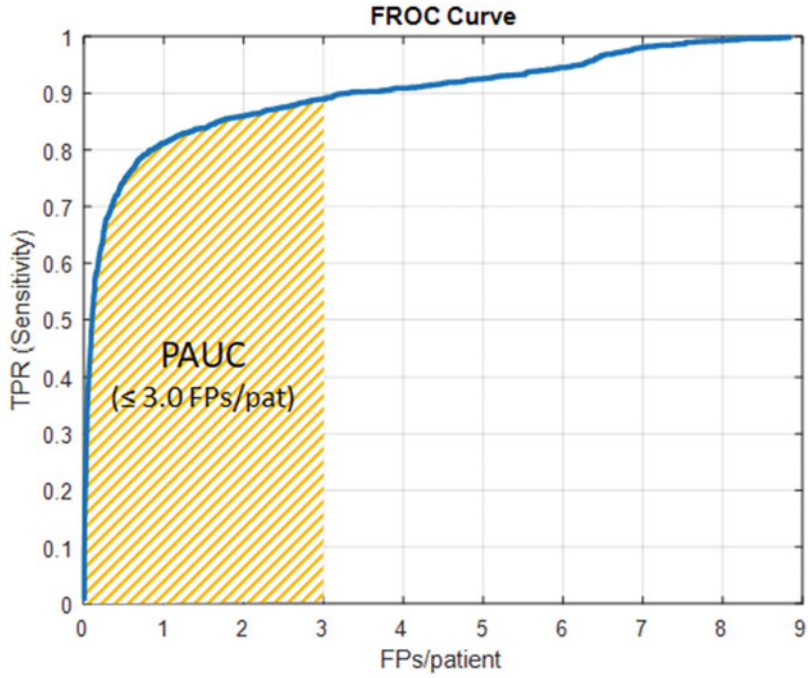

Fig. 4 Plot of an FROC curve for an ML detection algorithm. The FROC curve (blue line) shows the trade-off between object detection sensitivity (TPR) and the number of FPs per patient for all possible operating points. The full area under the FROC curve is not well defined, but a partial area may facilitate comparisons across algorithms. The figure shows a PAUC (shaded region) for ≤3.0 FPs/patient. However, AFROC-based summary metrics are more commonly used for characterizing/comparing FROC performance

The area under the alternate FROC (AFROC) curve with a jackknife method (JAFROC) was developed to provide confidence interval estimates and facilitate statistical performance comparisons across algorithms [56, 61]. AFROC provides an alternative way to summarize FROC data where the fraction of negative images falsely called positive are computed based on the highest FP score for each image in the dataset [58]. In this way, the unlimited x-axis of FROC curves is now bounded at 1 as shown in Fig. 5, and the area under the curve is well defined. Chakraborty's jackknife FROC (JAFROC) metric is the area under this AFROC calculated using a jackknife approach [56, 61].

Another common aggregate assessment for ML detection performance is the precision–recall (P–R) curve (*see* Fig. 6) which plots the trade-off between precision and recall across the full range of ML detection algorithm confidence scores [63].

As a reminder, precision (PPV) is a measure of how well the ML detection algorithm identifies only relevant abnormalities, while recall (TPR) is a measure of how well the algorithm finds all abnormalities. A better ML detection algorithm will have a higher precision at a fixed recall. Therefore, a larger area under the P–R

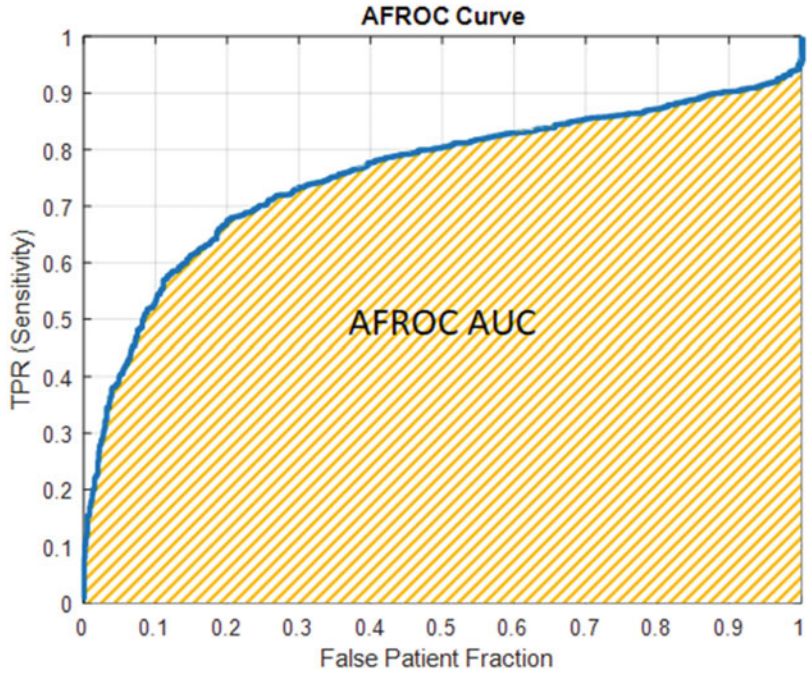

Fig. 5 Plot of an AFROC curve for the same ML detection algorithm given in Fig. 4. The AFROC curve (blue line) shows the trade-off between sensitivity (TPR) and the false patient fraction (fraction of patients with at least one FP) for all possible operating points. The area under the AFROC curve (shaded region) is often used to facilitate comparisons across object detection algorithms

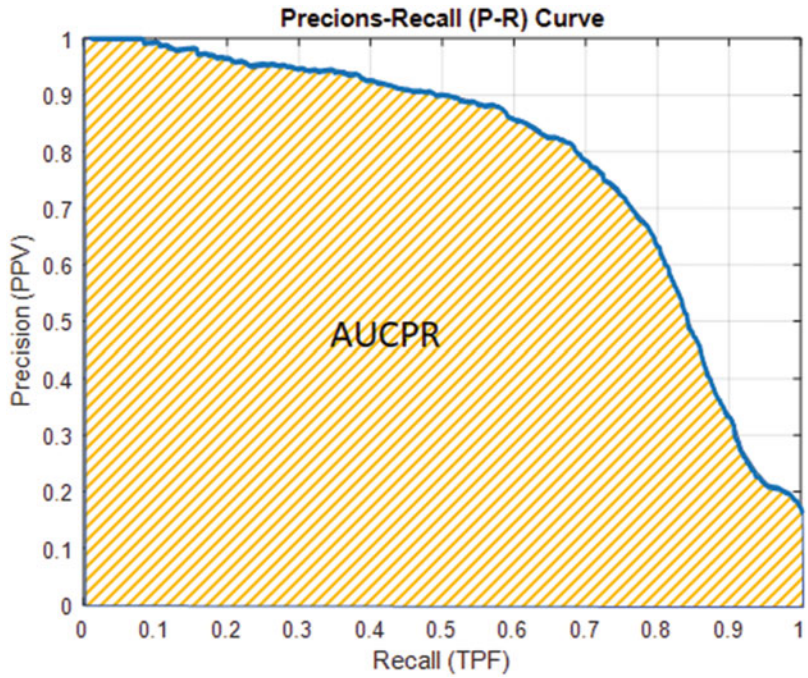

Fig. 6 Plot of a P–R curve for the same ML object detection algorithm as in Figs. 4 and 5. The P–R curve shows the trade-off in precision (PPV) as a function of recall (TPF). The area under the P–R curve (AUCPR) is an aggregate summary metric, for characterizing and comparing P–R curves across object detection algorithms

curve indicates improved performance compared to a competing algorithm, at least when the two P–R curves do not cross. The area under the P–R curve (AUCPR) is again an aggregate summary metric with the average precision (AP) as one estimation method developed in the information retrieval literature and has been used as a performance metric in ML Grand Challenges assessing ML localization algorithms [64, 65]. Nonparametric P–R curve and AP are commonly reported with one definition of AP given below.

- Average precision

$$AP = \sum_{n=1}^{N} (R_{n+1} - R_n) P_{\text{interp}}(R_{n+1}), \text{where } P_{\text{interp}}(R_{n+1}) = \max_{\widehat{R}:\widehat{R} \geq R_{n+1}} P\left(\widehat{R}\right)$$

Another approach is to use an 11-point interpolation by averaging the maximum precision for a set of 11 equally spaced recall levels [0, 0.1, 0.2, ..., 1] [63]. Parametric [66] and semi-parametric [67] methods for fitting the P–R curve and methods for estimating the AUCPR (e.g., trapezoidal estimators, interpolation estimators) have also been reported in the literature.

One of the complications in assessing an ML algorithm for abnormality detection is the need for determining a "correct" detection based on either an overlap measure for a bounding box/segmentation output or a distance metric for a marker output. Since ML algorithm performance depends on the "correct" detection criterion defined by an empirically chosen overlapping or distance parameter, a sensitivity analysis of the standalone performance across a range of overlap parameters is helpful to confirm that the performance estimate is reasonably stable or to at least understand how the choice of the criterion impacts performance. Moreover, while we have concentrated on detecting a single abnormality here, the abnormality detection metric discussed above can be generalized to multiple-object detection problems by reporting overall performance or assessing performance individually for each type of abnormality and averaging across abnormality types.

6.4 Triage Assessment

A triage ML algorithm analyzes images for findings suggestive of a target clinical condition, but instead of making a diagnosis or detection on the image, the algorithm is limited to generating a notification in the reading worklist or communicating directly to a specialist that a patient has a potential time-sensitive condition. Triage ML devices are often called computer-assisted triage and notification (CADt) devices. CADt is designed to allow a full clinical review earlier in the workflow than without the ML notification, given a true-positive (TP) finding by the algorithm. This can

benefit patients for conditions that are time critical by providing more timely care. For example, in cases involving suspected large vessel occlusion (LVO) stroke, a notification from an effective CADt device could allow a neuro-interventionalist to expeditiously treat the clot, potentially reducing some associated morbidity and mortality.

Another situation that CADt devices are useful is in a busy clinical environment where a large number of cases are queued waiting for clinician review. Instead of reading the cases in a first-in-first-out (FIFO) fashion, the clinician can review CADt flagged cases before non-flagged cases, thereby reducing the waiting time of the diseased patients.

In both situations described above, the sensitivity of the CADt for the target condition is critical so that the truly diseased patients benefit from earlier diagnosis and treatment. However, specificity is also important for the following reasons. An ML algorithm is unlikely to have 100% sensitivity, i.e., there are inevitably false-negative patients in the queue. These patients may be significantly delayed compared with FIFO reading if the triage algorithm has a large false-positive rate (i.e., low specificity). Moreover, too many false alarms may lower the vigilance of a specialist which in turn may affect their performance on the true-positive patients. Therefore, the metric sensitivity and specificity should be used as a pair to assess CADt performance. In the same spirit, the overall capability of the ML algorithm in distinguishing between patients with the condition and those without can be assessed via ROC analysis and the area under the ROC curve.

Despite its usefulness in evaluating a CADt device, the (sensitivity, specificity) pair and ROC performance are metrics of *diagnosis* and, at best, indirect measures of the true clinical effectiveness of an ML triage, i.e., reduction of the waiting time for patients with the target time-sensitive condition. Quantitative assessment of the clinical effectiveness of CADt devices in accelerating the review of patient images with the condition of concern is an open question. Among the efforts we are aware of, Thompson et al. [68] are developing an analytical approach based on the queueing theory to quantify the wait-time-saving of CADt. Under a clinical workflow model parameterized by disease prevalence, patient arrival rate, radiologist service rate, and number of radiologists on-site, their method allows computation of the average waiting time saved for a truly diseased patient due to the use of the CADt device where CADt performance is characterized by its sensitivity and specificity in diagnosing the condition of interest. This approach can potentially be useful in assessing the clinical effectiveness of CADt algorithms but requires further development and validation. Likewise, alternate approaches for assessing true CADt effectiveness in a clinical setting should be an area of continued research.

6.5 Utility of Standalone Performance Assessment

As mentioned previously, ML standalone assessment is primarily used to benchmark algorithm performance and compare with other ML algorithms or prior versions of the same algorithm to determine a performance change. Once a standalone dataset has been established and referenced, and the various performance metrics and criteria set, standalone testing can generally be applied in an efficient manner. Therefore, standalone testing is an important tool for assessing the potential bias in an ML algorithm. When a large diverse standalone dataset containing a range of patients with various demographic characteristics, a wide range of disease conditions, and the full range of acquisition technologies and protocols is available, ML performance can be estimated and compared both overall on the full dataset and in separate subgroups within this larger population to help identify where the ML may perform better and worse.

The standalone testing is also a critical tool for confirming a potential bias or disparity when this disparity is hypothesized, through specifically targeting the assessment to that subgroup of interest. Through standalone testing, ML performance can quickly be evaluated on the specific subgroup to determine if concern is warranted. The data requirements for this type of focused subgroup assessments may not need to be unusually large if the goal is to identify large disparities in performance when the ML algorithm is suspected to be performing poorly. Obviously, identifying more nuanced differences in performance across subgroups requires larger datasets.

Finally, standalone testing is a great tool for comparing ML algorithms. Again, it is ideal to obtain a large diverse real-world dataset to fully assess and benchmark an ML algorithm, but comparison can often be performed on much smaller enriched datasets where the main goal is to determine which algorithm or modification performs best, especially in the developmental phases of an algorithm's life cycle.

6.6 Modifications and Continuous Learning

One of the potential advantages of ML is its ability to quickly learn from new data such that it can remain current to changing patient demographics, clinical practice, and image acquisition technologies. This ability may result in large numbers of updates to an ML algorithm after it becomes available for clinical use. However, each update requires a systematic assessment. Modifications can range from infrequent algorithm updates all the way to continuously learning ML that adapts or learns from real-world experience/data on a continuous basis. This presents a challenge to both ML developers and regulatory bodies such as the FDA.

FDA's traditional paradigm of regulating ML devices is not designed for adaptive technologies, which adapt and optimize performance on a rapid timescale. With this in mind, the FDA is exploring a new, total product life cycle (TPLC) regulatory

approach that may potentially accommodate the rapid modification cycle of ML algorithms allowing for their efficient improvement and adaptation to the changing clinical environment while still providing effective safeguards that meet FDA's statutory requirements to ensure safety and effectiveness. To this end, the FDA released a proposed regulatory framework for modifications to AI/ML software as a medical device (SaMD) in 2019 as a discussion paper [69] requesting feedback from the public on the proposed framework. The proposed TPLC approach is based on [69]:

- The assurance of quality systems and good machine learning practices (GMLP).
- An initial premarket assurance of safety and effectiveness.
- A limited set of SaMD pre-specifications.
- A well-defined algorithm change protocol.

The algorithm change protocol is defined as the specific methods that will be used to achieve and appropriately control the risks of the SaMD pre-specifications [69].

This proposed framework is still under development, but the FDA did provide more details on their potential approach with the release of the AI/ML SaMD Action Plan in January 2021. The Action Plan was developed in response to the stakeholder feedback received on the proposed framework and to support innovative work in the regulation of medical device software and other digital health technologies.[9]

In response to the FDA's proposed framework, Feng et al. have been working to frame an AI/ML algorithm change protocol as an online statistical hypothesis testing problem [70]. The goal of their work was to investigate how "biocreep" resulting from repeated testing and adoption of modifications might lead to a gradual deterioration in ML performance. Feng et al. were able to show that biocreep would regularly occur when using policies with no error-rate guarantees but policies that included error-rate control were able to control biocreep without substantially impacting the ability to approve beneficial modifications [70]. This was an in-depth study of a very limited scope of potential ML modification problems as indicated by Feng et al. [70], and there remains a great deal of work to address the challenges around other types of modifications and conditions. The scientific community, especially interdisciplinary teams of clinicians, statisticians, and domain experts, are encouraged to take on this interesting and complex ML problem [71].

[9] https://www.fda.gov/media/106331/download

7 Clinical Performance Assessment with a Reader Study

Simply put, a reader study for the assessment of ML algorithms is to put the algorithm in the hands of clinicians and study the effectiveness of the algorithm in aiding the clinician's decision-making. In this chapter, a reader study generally refers to a study in which readers (e.g., radiologists) review and interpret medical images for a specified clinical task (e.g., diagnosis) and provide objective quantitative interpretation such as a rating of the likelihood that a condition is present. This is fundamentally different from a survey or questionnaire for the radiologist to indicate if they "like" the functionalities of the ML algorithm, which is not task-specific or particularly subjective (i.e., "beauty test"). Moreover, reader studies for ML in medical imaging typically consist of two arms: reading images without the ML algorithm and with the algorithm output for medical decision-making, thereby enabling a comparison of the reader's performance between with and without the ML aid.

It is fundamentally important to distinguish between *fixed-reader* study and *random-reader* study. When readers are treated as fixed and patient cases are treated as random samples from the patient population, the variability/uncertainty of the performance estimate (without ML or with ML) arises only from the random sample of patient cases. What does this mean? Let us assume we have a radiologist whose name is Barbara in a fixed-reader study and her true diagnostic performance over the entire patient population is A_B. In one experiment, the estimate of Barbara's diagnostic performance is $\overline{A}_B$ with a 95% confidence interval (CI) $\left[L_{\overline{A}_B}, U_{\overline{A}_B}\right]$. This means that if the experiment were repeated infinite number of times, *each time with Barbara reading images of a random sample of patients*, then the average of estimates $\overline{A}_B$ in these repeated experiments would be A_B, and the true value A_B would be within the estimated confidence intervals 95% of the time. In this sense, we say the performance estimate "$\overline{A}_B \left[L_{\overline{A}_B}, U_{\overline{A}_B}\right]$" *of radiologist Barbara* is generalizable to the patient population. Notice that this conclusion is only about Barbara but nobody else.

On the other hand, in a random-reader study where both readers and cases are treated as random effects, the population parameter of interest A is the (average) performance of the reader population over the population of patients. The variability/uncertainty of the performance estimate $\overline{A}$ in one experiment $\left[L_{\overline{A}}, U_{\overline{A}}\right]$ should account for both the randomness of readers and that of cases—which is not a trivial task (see next paragraph for relevant literature). The interpretation of such estimates is that, if the experiments were repeated infinite number of times, *each time with a random sample of readers reading a random sample of cases'*

images, then the average of the performance estimates $\overline{A}$ in these repeated experiments would be A, and the population performance A would be within the estimated CIs 95% of the time. In this sense, we say the performance estimate "$\overline{A}\ [L_{\overline{A}}, U_{\overline{A}}]$" generalizes to both the reader population and the patient population, i.e., the performance estimate represents the expected performance of a random reader reading a random case using a medical device (e.g., an ML algorithm). To distinguish from a fixed-reader study, a random-reader study is often referred to as a multi-reader multi-case (MRMC) study. As a passing note, this discussion also indicates that it is critical to specify the intended patient population and user population of a device so that a study can be designed to collect data from those populations.

The statistical methodology for generalizing the performance of an imaging device to both the population of readers and the population of cases was first developed by Dorfman, Berbaum, and Metz (DBM) [72]. Since then, many methodologies have been developed for the analysis of MRMC data such as the Obuchowski and Rockette (OR) [73] model based on a correlated ANOVA model; the bootstrap method by Beiden, Wagner, and Campbell [74]; and the U statistic method by Gallas [75]. Relationships among these methods have also been investigated [76, 77]. These early developments of MRMC analysis methods have focused on the area under the ROC curve (AUC) as a performance metric; some of these methods (e.g., OR and U statistic methods) have been extended to binary performance metrics [78], and all these methods have been validated with simulation studies [79] [80]. Some of these methods also have publicly available software tools, such as the integrated and updated OR–DBM method[10] and the U statistic method.[11]

The most widely used MRMC study design for comparing two modalities (e.g., without ML versus with ML) is the fully crossed (FC) design, in which every reader reads every case in both modalities. The advantage of pairing both readers and cases across two modalities is that it builds a positive correlation between the performance estimates of the two modalities, thereby reducing the variability of the performance difference and enhancing the power of detecting the performance difference. This reduction of variability can be easily appreciated by a simple formula

[10] Software | Medical Image Perception Laboratory Department of Radiology (uiowa.edu): https://perception.lab.uiowa.edu/software-0

[11] iMRMC: Software to do multi-reader multi-case analysis of reader studies: https://github.com/DIDSR/iMRMC

$$\mathrm{Var}\left[\widehat{A}_1 - \widehat{A}_2\right] = \mathrm{Var}\left[\widehat{A}_1\right] + \mathrm{Var}\left[\widehat{A}_2\right] - 2\rho\sqrt{\mathrm{Var}\left[\widehat{A}_1\right]\mathrm{Var}\left[\widehat{A}_2\right]},$$

where "Var" denotes variance; $\widehat{A}_1$ and $\widehat{A}_2$ are performance estimates for, e.g., without ML and with ML, respectively; and ρ is the correlation between $\widehat{A}_1$ and $\widehat{A}_2$ and is positive under normal circumstances. Pairing cases from two modalities sometimes is not advised due to safety concerns, for example, if both imaging modalities involve ionizing radiation to patients, imaging the patient twice may raise dose concerns. Fortunately, this is not generally an issue for the assessment of ML algorithms, and pairing cases in a "without ML versus with ML" comparison is feasible in many diagnostic situations.

The FC design has been regarded as the most powerful design in the sense that it makes full use of available readers and cases in collection of information. However, practically the workload of a radiologist may be limited, and oftentimes an investigator may have more cases than what readers can afford to read. Moreover, as multi-site evaluation becomes popular for better generalizability, the transfer of cases among different clinical sites can be logistically demanding. To overcome these limitations, Obuchowski [81] investigated the split-plot design, where different groups of readers read different groups of cases. The combined reader/case group can still be paired across modalities to reduce the variability of performance difference. Figure 7 provides a visual illustration of the FC design and the paired split-plot (PSP) design. What might be surprising is that the PSP design can be more powerful than the FC design, as shown by Hillis et al. with empirical data [82] and Chen et al. with both theoretical analysis and real-world data [83]. This may sound like a paradox since the FC design is regarded as "the most powerful design," but it is not. Referring to Fig. 7, suppose we have a certain number of readers and each of them can read the same number of cases. In the FC design, all the readers read the same cases (*see* Fig. 7, top), whereas, in the PSP design, readers are partitioned into two groups with each group reading the same number of cases from two different case sets (*see* Fig. 7, bottom). As such, the two designs involve the same amount of workload (i.e., number of image interpretations). However, the PSP design has reduced variability in performance estimates and performance difference estimates and hence increased statistical power, as proved by Chen et al. [83] because of the inclusion of additional cases. One way to understand this is that, with the same workload, reading difference cases (by half of the readers) gains more information than reading the same cases. This is also consistent with a common statistical sense: when we have more cases, the variability of the "mean" measured on the cases is reduced. In summary, the FC design is the most powerful *given the same*

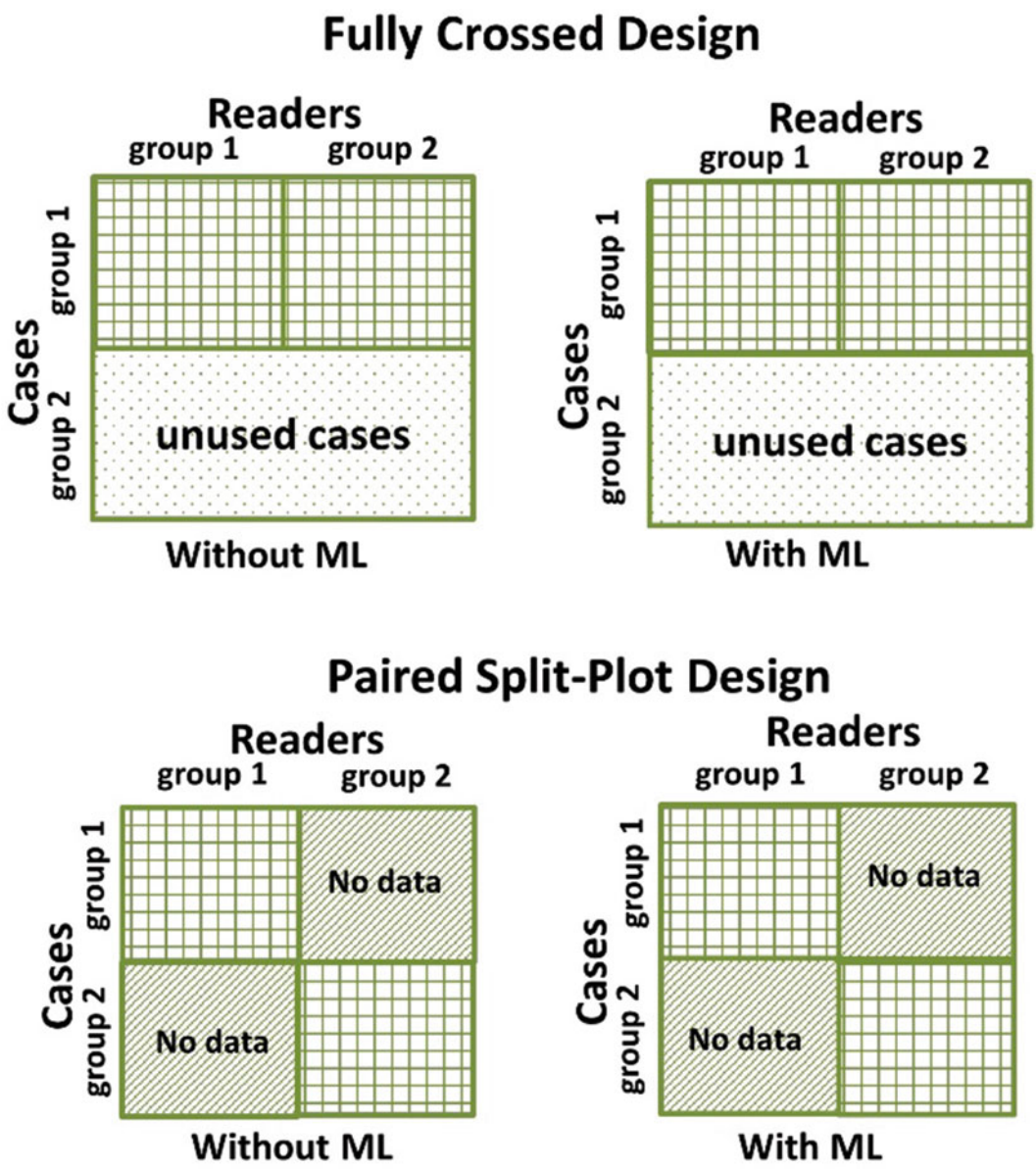

Fig. 7 Illustration of the fully crossed design and the paired split-plot design. The squares with grid can be understood as the data matrix collected in the reader study with each row representing a case, each column representing a reader, and each data element representing the rating of the case by the reader

number of readers and cases, but the PSP design can be more powerful *given the same number of image interpretations* with a price of collecting extra cases [83].

The design of an MRMC reader study involves a great deal of considerations including patient data collection (*see* Subheading 3), establishment of a reference standard (*see* Subheading 5), and many other aspects such as the recruitment and training of readers and

reading session design, e.g., *sequential reading*, where the readers read images with ML turned on immediately after reading without ML, or *concurrent reading*, where readers read images with ML turned on from the very beginning and this is typically compared against readers' performance reading images without ML in a separate session. It is worth noting that the discussion of the performance testing here is generally based on ML systems that are intended to "aid" or interface with an expert radiologist. The intended use of a model may warrant additional testing considerations related to human factors and human interpretability depending on how the model is integrated into the clinical workflow. Moreover, MRMC studies for the assessment of ML in imaging are often retrospective and controlled "laboratory" studies, in which typically only information related to the device of interest is presented to the readers (e.g., "image only" versus "image plus ML output"), whereas in real-world clinical practice, more information is often available to the physician, e.g., patient history, clinical tests, and/or other types of imaging exams. The diseased cases are often enriched when the natural prevalence is low in controlled laboratory studies. The purpose of such designs is to remove certain confounders and increase the statistical power to study the impact of the ML algorithm itself rather than the "absolute" performance of clinicians in the real world (as discussed in Subheading 3). However, consideration should still be given to ensure the study execution is as close to the clinical environment as possible and identify/mitigate potential biases, for example, the readers should be trained to use the ML algorithm as if they were instructed in the clinic. 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 [84], a consensus paper by Gallas et al. [8], as well as a tutorial paper by Wagner et al. [55].

8 Statistical Analysis

The statistical analysis plays a critical role in the assessment of ML performance but may be under-appreciated by many ML developers. For example, there are still publications that present point estimates of ML performance without quantification of uncertainties (standard deviations, confidence intervals). Even if uncertainty estimates are provided, the methods of uncertainty estimation are sometimes unclear or even inappropriate. Another example is the re-use of test data. One may follow the good practice of using independent datasets for ML training and testing. However, if the test data is repeatedly used, the seemingly innocent good practice may introduce optimistic bias to the performance estimate or even lead to a spurious discovery because the repeatedly measured

performance on the test dataset may inform training of the algorithm to adaptively fit the test data [85, 86]. As the quote goes, "if you torture the data long enough, it will confess." The lesson is, without following appropriate statistical principles, ML developers may be led to a blind alley due to statistical pitfalls: comparisons are made without statistical rigor, conclusions are drawn without appropriate data to substantiate, and spurious findings out of overfitting are celebrated. Statistical practices have a major impact on the ability to conduct reproducible research.

A good practice to avoid such pitfalls is, for any performance assessment study—either standalone performance assessment or an MRMC study—to *pre-specify* a statistical analysis plan (SAP) with valid statistical methods. The word "pre-specify" is emphasized because post hoc analyses can inflate the experiment-wise type I error rate and endanger the scientific validity of an otherwise well-designed and well-conducted study. Below we list exemplar elements in an SAP for ML development and assessment. We note that not all of them are necessarily applicable to a specific study. A specific SAP should be consistent with the study objectives, designs, nature of data, and statistical analysis methods.

1. Primary hypotheses and secondary hypotheses that are consistent with the primary and secondary goals of a study. This also necessarily involves choosing appropriate performance metrics (*see* Subheading 6 for different metrics corresponding to different clinical tasks). For example, the primary goal of an MRMC study might be to show the radiologists using an ML algorithm perform significantly better than without using the algorithm in the task of distinguishing between benign and malignant brain tumors, and a secondary goal might be to show the radiologist using an ML algorithm has significantly better specificity (S_p) at a given sensitivity. Then the null (H_0) and alternative (H_1) primary hypotheses can be stated as

$$H_0 : \mathrm{AUC}_{\text{with ML}} = \mathrm{AUC}_{\text{without ML}}; H_1 : \mathrm{AUC}_{\text{with ML}} > \mathrm{AUC}_{\text{without ML}}.$$

And the secondary hypotheses may be stated as

$$H_0 : S_{p_{\text{with ML}}} = S_{p_{\text{without ML}}}; H_1 : S_{p_{\text{with ML}}} > S_{p_{\text{without ML}}}.$$

2. A plan for use of patient data in various stages of ML algorithm development and performance assessment. As discussed in Subheading 3, patient data are used in both the development and assessment of ML algorithms. A pre-specified plan for appropriate use of patient data is crucial for achieving the goals of algorithm development and performance validation and controlling various sources of bias in the process.

3. Methods for analyzing the study data to estimate the pre-specified performance metric, the uncertainty of the metric (e.g., standard deviations and confidence intervals) accounting for all sources of variability including reference standard as needed, and the test statistic for hypothesis testing. It is critically important to examine if the assumptions behind the statistical methods are appropriate for the data and, when necessary, use an alternative method to verify the results.
4. Sample size determination. In a standalone performance assessment study, this is to determine the number of patients to be included in the study such that the study data are representative of the intended patient population (*see* Subheading 3.3.3) and, when applicable, the study has sufficient statistical power (typically set to be >80%) to detect a significant effect (e.g., superior performance compared with a control). With a single source of variability, standard statistical methods and software tools are often useful for sizing a standalone performance assessment study.

   In an MRMC study, both the number of readers and the number of cases need to be determined. Sample size determination is again mainly for assuring a reasonable chance of success in the study planning stage. From a technical point of view (i.e., not taking into consideration practical issues such as budget), sample size is typically determined by considering (1) that the sample sizes are large enough to include samples that represent the intended patient and reader populations and (2) the sample sizes are sufficient to achieve a target statistical power in a hypothesis testing study. Due to the complexity, specialized software tools can be used for sizing an MRMC study [87], and the MRMC software tools cited in Subheading 7 provide the sizing functionality.

   The information needed for sizing a pivotal study is often obtained in a pilot study, as discussed in Subheading 3. However, sometimes the pilot study is too limited to provide reliable information, and one may find attempting to re-size a pivotal study after an interim analysis of the data. Naively re-sizing the study based on information obtained in the same study may inflate the type I error rate. Huang et al. [88] developed an approach that allows adaptive re-sizing of an MRMC study with information obtained in an interim analysis such that the statistical power is adjusted to a target value and the type I error rate is retained by paying a statistical penalty in the final hypothesis testing.
5. A plan for adjusting p-values and/or confidence intervals for multiple comparisons or hypothesis tests.
6. A plan for handling missing data and assessing the impact of missing data (e.g., missing reader data, missing follow-up data

to confirm negative results) on the study conclusions. Although statistical techniques may be used to address issues of loss-to-follow-up and missing data, these techniques often employ major assumptions that cannot be fully validated for a particular study. Therefore, the best way to address issues of missing data due to loss-to-follow-up is to plan to minimize its occurrence during the planning and management of the clinical study. Nevertheless, the study protocol should pre-specify appropriate statistical data analysis methods, in addition to sensitivity analyses, for handling missing data.

9 Summary Remarks

In this chapter, we provided an overview of a performance assessment framework for imaging-based ML algorithms. We discussed general considerations in study design and data collection, establishment of a reference standard, algorithm documentation, algorithm standalone performance as well as clinical reader studies, and statistical data analysis in performance testing. We believe that these topics are relevant not only in the regulatory setting but also to reproducible science and technology development. Because patient data and clinical experts' annotations are used in both the *development* and *assessment* of ML algorithms, performance assessment should be considered from the very beginning of development to make efficient use of available data. In addition, performance assessment and algorithm development (e.g., tuning, internal validation) are often iterative; meaningful assessment methodologies and tools are not only meant to make the assessment *rigorous* but also *cost-effective*. Furthermore, performance assessment methodologies are also tremendously helpful to assure quality and reproducibility, control bias, and avoid pitfalls and blind alleys.

Machine learning technologies are still rapidly evolving, and their applications in medicine and brain imaging in particular are expanding. It is widely recognized that ML is playing a pivotal role in revolutionizing medicine and promoting public health to a new level. Accompanying these potential developments are new research questions on assessment methodologies. We have touched upon topics in this chapter such as novel types of clinical applications enabled by ML and continuous learning ML. Other exciting topics may include improvement and assessment of robustness and generalizability of ML algorithms, synthetic data augmentation, characterization of bias/fairness, and uncertainty-aware ML algorithms that output not only clinical conditions of interest but also "I don't know," among many others. We believe that assessment methodologies and regulatory science play a critical role in fully realizing the great potential of ML in medicine, in facilitating ML device innovation, and in accelerating the translation of these technologies from bench to bedside to the benefit of patients.

Acknowledgments

The authors thank Drs. Robert Ochs, Alexej Gossmann, and Aldo Badano for carefully reviewing the manuscript and providing helpful comments.

References

1. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML (2019) Deep learning in medical imaging and radiation therapy. Med Phys 46(1):e1–e36. https://doi.org/10.1002/mp.13264
2. Lui YW, Chang PD, Zaharchuk G, Barboriak DP, Flanders AE, Wintermark M, Hess CP, Filippi CG (2020) Artificial intelligence in neuroradiology: current status and future directions. Am J Neuroradiol 41(8):E52–E59. https://doi.org/10.3174/ajnr.A6681
3. U.S. Food and Drug Administration (2017) De Novo classification process (Evaluation of Automatic Class III Designation). Guidance for Industry and Food and Drug Administration Staff
4. U.S. Food and Drug Administration (2014) The 510(k) program: evaluating substantial equivalence in premarket notifications [510 (k)]. Guidance for Industry and Food and Drug Administration Staff
5. U.S. Food and Drug Administration (2012) Factors to consider when making benefit-risk determinations in medical device premarket approval and De Novo classifications. Guidance for Industry and Food and Drug Administration Staff
6. U.S. Food and Drug Administration (2018) Benefit-risk factors to consider when determining Substantial equivalence in premarket notifications (510(k)) with different technological characteristics. Guidance for Industry and Food and Drug Administration Staff
7. U.S. Food and Drug Administration (2021) Requests for feedback and meetings for medical device submissions: the Q-submission program. Guidance for Industry and Food and Drug Administration Staff
8. Gallas BD, Chan HP, D'Orsi CJ, Dodd LE, Giger ML, Gur D, Krupinski EA, Metz CE, Myers KJ, Obuchowski NA, Sahiner B, Toledano AY, Zuley ML (2012) Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. Acad Radiol 19(4):463–477. https://doi.org/10.1016/j.acra.2011.12.016
9. Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning. Series in statistics, 2nd (corrected 12th printing) edn. Springer, New York
10. Chan H-P, Sahiner B, Wagner RF, Petrick N (1999) Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys 26(12):2654–2668
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster R (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410. https://doi.org/10.1001/jama.2016.17216
12. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, New York
13. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 162(1):W1–W73. https://doi.org/10.7326/m14-0698
14. Du B, Wang Z, Zhang L, Zhang L, Liu W, Shen J, Tao D (2019) Exploring representativeness and informativeness for active learning. arXiv:1904.06685
15. Huang S, Jin R, Zhou Z (2014) Active learning by querying informative and representative examples. IEEE Trans Pattern Anal Mach Intell 36:1936–1949
16. Sharma D, Shanis Z, Reddy CK, Gerber S, Enquobahrie A (2019) Active learning technique for multimodal brain tumor segmentation using limited labeled images. In: Wang Q, Milletari F, Nguyen HV et al (eds) Domain adaptation and representation transfer and medical image learning with less labels and imperfect data. Springer International Publishing, Cham, pp 148–156

17. Hao R, Namdar K, Liu L, Khalvati F (2021) A transfer learning–based active learning framework for brain tumor classification. Front Artif Intell 4(61):635766. https://doi.org/10.3389/frai.2021.635766
18. Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (2009) Dataset shift in machine learning. MIT Press, Cambridge MA
19. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recogn 45(1):521–530. https://doi.org/10.1016/j.patcog.2011.06.019
20. Storkey A (2009) When training and test sets are different: characterizing learning transfer. In: Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (eds) Dataset shift in machine learning. MIT Press, Cambridge, MA, pp 3–28
21. Goldenberg I, Webb G (2019) Survey of distance measures for quantifying concept drift and shift in numeric data. Knowl Inf Syst 60:591–615. https://doi.org/10.1007/s10115-018-1257-z
22. Rabanser S, Günnemann S, Lipton ZC (2018) Failing loudly: an empirical study of methods for detecting dataset shift. arXiv:1810.11953
23. Dockès J, Varoquaux G, Poline J-B (2021) Preventing dataset shift from breaking machine-learning biomarkers. arXiv:2107.09947
24. Turhan B (2012) On the dataset shift problem in software engineering prediction models. Empir Softw Eng 17(1):62–74. https://doi.org/10.1007/s10664-011-9182-8
25. U.S. Food and Drug Administration (2007) Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests. vol 2007. U.S Food and Drug Administration, Silver Spring
26. Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley
27. Suresh H, Guttag JV (2021) A framework for understanding sources of harm throughout the machine learning life cycle. arXiv:190110002 [cs, stat]
28. Hooker S (2021) Moving beyond "algorithmic bias is a data problem". Patterns 2(4):100241. https://doi.org/10.1016/j.patter.2021.100241
29. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, Shah N, Sung L (2021) Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Appl Clin Inform 12(4):808–815. https://doi.org/10.1055/s-0041-1735184
30. National Academies of Sciences E, Medicine (2019) Reproducibility and replicability in science. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303
31. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V (2009) Repeatability of published microarray gene expression analyses. Nat Genet 41(2):149–155. https://doi.org/10.1038/ng.295
32. Baggerly KA, Coombes KR (2009) Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. Ann Appl Stat 3(4):1309–1334, 1326
33. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, Fox E, Larochelle H (2020) Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). arXiv:2003.12206
34. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 26(9):1320–1324. https://doi.org/10.1038/s41591-020-1041-y
35. Mongan J, Moy L, Charles E, Kahn J (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2(2):e200029. https://doi.org/10.1148/ryai.2020200029
36. El Naqa I, Boone JM, Benedict SH, Goodsitt MM, Chan HP, Drukker K, Hadjiiski L, Ruan D, Sahiner B (2021) AI in medical physics: guidelines for publication. Med Phys 48(9):4711–4714. https://doi.org/10.1002/mp.15170
37. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, Darzi A, Holmes C, Yau C, Moher D, Ashrafian H, Deeks JJ, Ferrante di Ruffano L, Faes L, Keane PA, Vollmer SJ, Lee AY, Jonas A, Esteva A, Beam AL, Panico MB, Lee CS, Haug C, Kelly CJ, Yau C, Mulrow C, Espinoza C, Fletcher J, Moher D, Paltoo D, Manna E, Price G, Collins GS, Harvey H, Matcham J, Monteiro J, ElZarrad MK, Ferrante di Ruffano L, Oakden-Rayner L, McCradden M, Keane PA, Savage R, Golub R, Sarkar R, Rowley S, The S-A, Group C-AW, Spirit AI, Group C-AS, Spirit

AI, Group C-AC (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26(9):1351–1363. https://doi.org/10.1038/s41591-020-1037-7

38. Collins G, Moons K (2019) Reporting of artificial intelligence prediction models. Lancet 393:1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6
39. U.S. Food and Drug Administration (2012) Clinical performance assessment: Considerations for computer-assisted detection devices applied to radiology images and radiology device data – premarket approval (PMA) and premarket notification [510(k)] submissions – Guidance for industry and FDA staff. https://www.fda.gov/media/77642/download. Accessed 31 Oct 2021
40. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23(7):903–921
41. Petrick N, Sahiner B, Armato SG III, Bert A, Correale L, Delsanto S, Freedman MT, Fryd D, Gur D, Hadjiiski L, Huo Z, Jiang Y, Morra L, Paquerault S, Raykar V, Salganicoff M, Samuelson F, Summers RM, Tourassi G, Yoshida H, Zheng B, Zhou C, Chan H-P (2013) Evaluation of computer-aided detection and diagnosis systems. Med Phys 40:087001–087017
42. Steyerberg EW (2019) Overfitting and optimism in prediction models. In: Clinical prediction models. Springer, pp 95–112
43. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ (2017) Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 30(4):449–459. https://doi.org/10.1007/s10278-017-9983-4
44. Zhang YJ (1996) A survey on evaluation methods for image segmentation. Pattern Recogn 29(8):1335–1346. https://doi.org/10.1016/0031-3203(95)00169-7
45. Zhang YJ (2001) A review of recent evaluation methods for image segmentation. In: Proceedings of the sixth international symposium on signal processing and its applications (Cat. No.01EX467), 13–16 Aug 2001. vol. 141, pp 148–151. https://doi.org/10.1109/ISSPA.2001.949797
46. Meyer CR, Johnson TD, McLennan G, Aberle DR, Kazerooni EA, Macmahon H, Mullan BF, Yankelevitz DF, van Beek EJR, Armato SG 3rd, McNitt-Gray MF, Reeves AP, Gur D, Henschke CI, Hoffman EA, Bland PH, Laderach G, Pais R, Qing D, Piker C, Guo J, Starkey A, Max D, Croft BY, Clarke LP (2006) Evaluation of lung MDCT nodule annotation across radiologists and methods. Acad Radiol 13(10):1254–1265
47. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 15(1):29. https://doi.org/10.1186/s12880-015-0068-x
48. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302. https://doi.org/10.2307/1932409
49. Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytol 11(2):37–50
50. Willem (2017) FI/Dice-Score vs IoU. Cross Validated. https://stats.stackexchange.com/questions/273537/f1-dice-score-vs-iou/276144#276144. Accessed 9/29/2021
51. Fenster A, Chiu B (2005) Evaluation of segmentation algorithms for medical imaging. In: 2005 IEEE engineering in medicine and biology 27th annual conference, 17–18 Jan 2006. pp 7186–7189. https://doi.org/10.1109/IEMBS.2005.1616166
52. Tharwat A (2021) Classification assessment methods. Appl Comput Inform 17(1):168–192. https://doi.org/10.1016/j.aci.2018.08.003
53. Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5(2):1
54. Obuchowski NA (2003) Receiver operating characteristic curves and their use in radiology. Radiology 229(1):3–8
55. Wagner RF, Metz CE, Campbell G (2007) Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol 14(6):723–748
56. Chakraborty DP (2018) Observer performance methods for diagnostic imaging: foundations, modeling, and applications with r-based examples. Imaging in medical diagnosis and therapy. CRC Press, Boca Raton, FL
57. ICRU (2008) Receiver operating characteristic analysis in medical imaging. Report 79. International Commission of Radiation Units and Measurements, Bethesda, MD
58. He X, Frey E (2009) ROC, LROC, FROC, AFROC: an alphabet soup. J Am Coll Radiol 6(9):652–655
59. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH (1977) A free response approach to the measurement and characterization of radiographic observer performance. Proc SPIE 127:124–135

60. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM (2002) Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 29(12):2861–2870. https://doi.org/10.1118/1.1524631
61. Chakraborty DP (2006) Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. Acad Radiol 13(10):1187–1193
62. Chakraborty DP (2006) A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol 51(14):3449–3462
63. Padilla R, Netto SL, Silva EABd (2020) A survey on performance metrics for object-detection algorithms. In: 2020 international conference on systems, signals and image processing (IWSSIP), 1–3 July 2020. pp 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130
64. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4
65. ImageNet (2017) ImageNet object localization challenge. Kaggle. https://www.kaggle.com/c/imagenet-object-localization-challenge/. Accessed 10/22/2021 2021
66. Liu Z, Bondell HD (2019) Binormal precision–recall curves for optimal classification of imbalanced data. Stat Biosci 11(1):141–161. https://doi.org/10.1007/s12561-019-09231-9
67. Sahiner B, Chen W, Pezeshk A, Petrick N (2016) Semi-parametric estimation of the area under the precision-recall curve. In: SPIE medical imaging. International Society for Optics and Photonics, pp 97870D-97870D-97877
68. Thompson E, Levine G, Chen W, Sahiner B, Li Q, Petrick N, Samuelson F (2022) Wait-time-saving analysis and clinical effectiveness of computer-aided triage and notification (CADt) devices based on queueing theory. In: Taylor-Phillips CRM-TaS (ed) Medical imaging 2022: Image perception, observer performance, and technology assessment, San Diego, CA, SPIE, p accepted
69. U.S. Food and Drug Administration (2019) 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. U.S Food and Drug Administration. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. Accessed 31 Oct 2021
70. Feng J, Emerson S, Simon N (2021) Approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep. Biometrics 77(1):31–44. https://doi.org/10.1111/biom.13379
71. Pennello G, Sahiner B, Gossmann A, Petrick N (2021) Discussion on "approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep" by Jean Feng, Scott Emerson, and Noah Simon. Biometrics 77(1):45–48. https://doi.org/10.1111/biom.13381
72. Dorfman DD, Berbaum KS, Metz CE (1992) Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Investig Radiol 27(9):723–731
73. Obuchowski NA, Rockette HE (1995) 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. https://doi.org/10.1080/03610919508813243
74. Beiden SV, Wagner RF, Campbell G (2000) Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad Radiol 7(5):341–349
75. Gallas BD (2006) One-shot estimate of MRMC variance: AUC. Acad Radiol 13(3):353–362
76. Hillis SL, Berbaum KS, Metz CE (2008) Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol 15(5):647–661
77. Gallas BD, Bandos A, Samuelson FW, Wagner RF (2009) A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. Commun Stat Theory Methods 38(15):2586–2603. https://doi.org/10.1080/03610920802610084
78. Gallas BD, Pennello GA, Myers KJ (2007) Multireader multicase variance analysis for binary data. J Opt Soc Am A 24(12):B70–B80
79. Metz CE (1995) The Dorfman/Berbaum/Metz method for testing the statistical significance of ROC differences: validation studies with continuously-distributed data. The Farwest image perception conference to be given October 13, 1995 in Philadelphia, PA
80. Chen W, Wunderlich A, Petrick N, Gallas BD (2014) Multireader multicase reader studies

with binary agreement data: simulation, analysis, validation, and sizing. J Med Imaging (Bellingham) 1(3):031011–031011. https://doi.org/10.1117/1.JMI.1.3.031011

81. Obuchowski NA (2009) Reducing the number of reader interpretations in MRMC studies. Acad Radiol 16(2):209–217
82. Obuchowski NA, Gallas BD, Hillis SL (2012) Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. Acad Radiol 19(12):1508–1517. https://doi.org/10.1016/j.acra.2012.09.012
83. Chen W, Gong Q, Gallas BD (2018) Paired split-plot designs of multireader multicase studies. J Med Imaging (Bellingham) 5(3):031410. https://doi.org/10.1117/1.JMI.5.3.031410
84. U.S. Food and Drug Administration (2020) 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
85. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A (2015) The reusable holdout: preserving validity in adaptive data analysis. Science 349(6248):636–638
86. Gossmann A, Pezeshk A, Wang Y-P, Sahiner B (2021) 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 3:692–714. https://doi.org/10.1137/20M1333110
87. Hillis SL, Obuchowski NA, Berbaum KS (2011) Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol 18(2):129–142
88. Huang Z, Samuelson F, Tcheuko L, Chen W (2020) Adaptive designs in multi-reader multicase clinical trials of imaging devices. Stat Methods Med Res 29(6):1592–1611. https://doi.org/10.1177/0962280219869370