AAPM Task Group Report 273: Recommendations on Best Practices for AI and Machine Learning for Computer-Aided Diagnosis in Medical Imaging

Lubomir Hadjiiski
Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA

Kenny Cha
U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Heang-Ping Chan
Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA

Karen Drukker
Department of Radiology, University of Chicago, Chicago, Illinois, USA

Lia Morra
Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy

Janne J. Näppi
3D Imaging Research, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

Berkman Sahiner
U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Hiroyuki Yoshida
3D Imaging Research, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

Quan Chen
Department of Radiation Medicine, University of Kentucky, Lexington, Kentucky, USA

Thomas M. Deserno
Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

Hayit Greenspan
Department of Biomedical Engineering, Faculty of Engineering, Tel Aviv University, Tel Aviv, Israel & Department of Radiology, Ichan School of Medicine, Mt Sinai, NYC, NY, USA

Henkjan Huisman
Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands

Zhimin Huo
Tencent America, Palo Alto, CA

Richard Mazurchuk
Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA

Nicholas Petrick
U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Daniele Regge
Radiology Unit, Candiolo Cancer Institute, FPO-IRCCS, Candiolo
Department of Surgical Sciences, University of Turin, Turin, Italy

Ravi Samala
U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Ronald M. Summers
Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, Maryland, USA

Kenji Suzuki
Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan

Georgia Tourassi
Oak Ridge National Lab, Oak Ridge, Tennessee, USA

Daniel Vergara
Department of Radiology, Yale New Haven Hospital, New Haven, Connecticut, USA

Samuel G. Armato, III
Department of Radiology, University of Chicago, Chicago, Illinois, USA

Disclosure Statement

The Chair of the AAPM Task Group 273 has reviewed the required Conflict of Interest statement on file for each member of AAPM Task Group 273 and determined that disclosure of potential Conflicts of Interest is an adequate management plan. Disclosures of potential

Conflicts of Interest for each member of AAPM Task Group 273 are found at the close of this document.

Running title

AAPM Task Group Report 273

Correspondence:

Lubomir Hadjiiski
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
MIB C476
Ann Arbor, MI 48109
Telephone: (734) 647-7428
Fax: (734) 615-5513
E-mail: lhadjisk@umich.edu

Table of Contents

Report of AAPM Task Group 273

The purpose of this report is to provide recommendations on best practices and standards for the development and performance assessment of computer-aided decision support systems at the time when machine learning techniques continue to evolve, and CAD applications expand to new stages of the patient care process. The various steps of development are covered, including (1) data collection, (2) establishing reference standards, (3) model development, (4) performance assessment, and (5) translation to clinical practice. The goal of the report is to emphasize the proper training and validation methods for machine learning algorithms that may improve their generalizability and reliability and accelerate the adoption of CAD-AI systems for clinical decision support.

Abstract

Rapid advances in artificial intelligence (AI) and machine learning, and specifically in deep learning (DL) techniques, have enabled broad application of these methods in health care. The promise of the DL approach has spurred further interest in computer-aided diagnosis (CAD) development and applications using both 'traditional' machine learning methods and newer DL-based methods. We use the term CAD-AI to refer to this expanded clinical decision support environment that uses traditional and DL-based AI methods.

Numerous studies have been published to date on the development of machine learning tools for computer-aided, or AI-assisted, clinical tasks. However, most of these machine learning models are not ready for clinical deployment. It is of paramount importance to ensure that a clinical decision support tool undergoes proper training and rigorous validation of its generalizability and robustness before adoption for patient care in the clinic.

To address these important issues, the American Association of Physicists in Medicine (AAPM) Computer-Aided Image Analysis Subcommittee (CADSC) is charged, in part, to develop recommendations on practices and standards for the development and performance assessment of computer-aided decision support systems. The committee has previously published two opinion papers on the evaluation of CAD systems and issues associated with user training and quality assurance of these systems in the clinic. With machine learning techniques continuing to evolve and CAD applications expanding to new stages of the patient care process, the current task group report considers the broader issues common to the development of most, if not all, CAD-AI applications and their translation from the bench to the clinic. The goal is to bring attention to the proper training and validation of machine

learning algorithms that may improve their generalizability and reliability and accelerate the adoption of CAD-AI systems for clinical decision support.

1 Introduction

We are witnessing extensive development and an explosion of applications based on deep learning (DL) or "artificial intelligence (AI)" technology across various fields in recent years. Many applications in robotics, transportation, surveillance, Internet, and popular games have achieved high degrees of success and raised unprecedented enthusiasm for AI. Rapid advances in machine learning, and specifically in DL techniques, have enabled broad application of these methods in health care. In medical imaging, computer-aided diagnosis (CAD) using traditional machine learning techniques was introduced into the clinic over two decades ago; however, traditional approaches that use manually designed image features (i.e., mathematical descriptors) and classifiers with small numbers of parameters may yield limited performance for some complex tasks. DL is a representation learning technique in which a multi-layer neural network with millions of interconnecting weights automatically learns relevant features and information from the input data and models the expected outcome guided by a large set of training samples. The increasing accessibility to low-cost computational power and data storage further enables the development of DL models. The promise of the DL approach has spurred a new era of development of CAD-AI applications for clinical decision support in various stages of the patient care process; we use the term CAD-AI to refer to this expanded clinical decision support environment that uses traditional and DL-based AI methods (Figure 1).

Numerous studies have been published to date on the development of machine learning tools for computer-aided, or AI-assisted, clinical tasks. In a recent review of publications related to machine learning-based detection and prognosis of COVID-19 using chest radiographs and CT scans, Roberts et al. [1] concluded that none of the models were of potential clinical use due to methodological flaws and/or underlying biases. In another review of the design, reporting standards, and claims of studies that compared the performance of the DL algorithms applied to medical images with that of expert clinicians, Nagendran et al. [2] concluded that only a few prospective DL studies and randomized trials had been performed and that the rest of the studies were at high risk for bias. In a systematic review on the diagnostic accuracy of DL algorithms, Aggarwal at al. [3] found high heterogeneity and extensive variation in methodology, terminology, and outcome measures among the studies, all of which could lead to an overestimation of the diagnostic accuracy of DL algorithms applied to medical images. In a review of over 500 studies that evaluated the performance of AI algorithms for diagnostic analysis of medical images, Kim et al. [4] reported that nearly all were designed as proof-of-concept technical feasibility studies and did not incorporate design features that are recommended for robust validation of the real-world clinical performance of AI algorithms. These reviews reveal that the majority of machine learning models developed to date seem to be far from ready for clinical deployment despite the reported levels of performance.

Regardless of the underlying machine learning methods used for development of CAD tools, it is of paramount importance to ensure that a clinical decision support tool has undergone proper training and rigorous validation of its generalizability and robustness before the adoption of such tools for patient care in the clinic. To address these important

issues, the American Association of Physicists in Medicine (AAPM) Computer-Aided Image Analysis Subcommittee (CADSC) is charged, in part, to develop recommendations on practices and standards for the development and performance assessment of computer-aided decision support systems. The CADSC has previously published two papers to convey the opinions of CADSC members on proper practices for the training, evaluation, and quality assurance of CAD systems [5, 6]. With machine learning techniques continuing to evolve and CAD applications expanding to new stages of the patient care process (Figure 1), this task group report addresses the broad issues common to the development of most, if not all, CAD-AI applications and their translation from the bench to the clinic. The various steps of development will be covered, including data collection, establishing reference standards, model development, performance assessment, and translation to clinical practice, as summarized in Figure 2. The goal is to bring attention to proper training and validation methods for machine learning algorithms that may improve their generalizability and reliability and accelerate the adoption of CAD-AI systems for clinical decision support.

2 Data

The most fundamental step for the development of a CAD-AI tool is to define the use case and the population to which the CAD-AI tool is to be applied. As a guiding principle, data collected for the training, validation, and testing of a CAD-AI tool should reflect the intended use case and population while at the same time allowing for the replication of results in a real-world clinical setting. It cannot be overemphasized that improper data collection practices may likely introduce bias and create a misleading perception of model performance, especially in subpopulations that may not be appropriately represented in the study dataset. In study reports, the data collection process must be described in detail to demonstrate scientific rigor and should include inclusion and exclusion criteria as well as the target patient demographics.

This section covers the topics of data collection (including case sampling, public databases, ethics, and quality considerations), data augmentation, and data harmonization. The topic of labels that might accompany collected data will be covered in the Reference Standards section (section 3).

2.1 Data Collection

2.1.1 Data collection and case sampling

System development with consecutively sampled cases from multiple sites over a defined period of image acquisition dates [7] is the best way to achieve replication of performance in a real-world clinical setting. In some machine learning applications for which the proportion of different case groups is highly imbalanced in the population, however, consecutive data collection is impractical, and the training dataset must be collected with methods such as stratified sampling to enrich some of the groups. For example, in the case of screening mammography, stratifying samples across the positive and negative groups is needed because the yield of malignancy is only 0.5%. **Stratified sampling** [8] splits the population into non-overlapping groups (or strata) and then samples within each strata to achieve the desired balance among different strata; if applied accurately, stratified sampling can enhance the generalizability of a model relative to training without stratification. In practice, many development studies are performed using a **convenience sample approach** [9], where cases that are conveniently available to the developers are the ones collected for the study.

Especially in new research areas, the availability of only a convenience sample should not prevent a study from going forward; however, claims about CAD-AI system performance in such studies should be made with utmost care to reflect the reality that the results are likely not generalizable.

Several recent studies have indicated that systems developed and tested with data from one collection site failed to achieve similar test results when applied to data from a different site [10-13]. For this reason, especially for validation studies, it is essential to have **multi-site data collection** [14, 15] and to assure that the data collection is diverse in terms of subject population, disease severity, vendor/imaging system, and image acquisition protocol. Development studies that use **single-site data collection** are essential for new advancements in a time-efficient manner, but strong limitations about the assessed performance should be acknowledged.

2.1.2 Public databases

In CAD-AI development, each research group typically uses its resources to collect its own database, which is likely to be smaller in number than desirable and lacking the real-world diversity of patient demographics and image acquisition parameters that exist across institutions. Furthermore, this isolation of databases prohibits the direct comparison of the performance of systems reported in the literature [16, 17].

Publicly available image databases overcome these shortcomings by providing a free, accessible resource for the international medical imaging research community. The creation of a **public database** is not as simple as depositing one or more existing local databases on a web site or crowd-sourcing the uploading of images and associated information. The nature of the public database should be prospectively determined in terms of the clinical task(s) it may be expected to address, the range of disease presentations to be represented by those cases, the associated metadata it will include, and the reference standard it will incorporate. The need for a quality assurance (QA) process for data in a public database cannot be overemphasized [18, 19]: adherence to the case inclusion/exclusion criteria, proper de-identification of protected health information (PHI), image quality, and reference standard integrity must all be verified before the database can be released for public access. In addition, the FAIR (Findable, Accessible, Interoperable and Reusable) principles must be followed to the extent possible in designing public datasets to assist both human users and their computational agents in the discovery of, access to, and integration and analysis of the data [20].

Public databases are resources of growing importance for the advancement of machine learning algorithms in medical imaging and clinical decision support in general. These databases play important roles in algorithm development, training/testing, validation, and performance assessment; in short, they expedite the ability of research groups to contribute to the field. Investigators who use these databases have an obligation to understand the limitations of the databases and to use them in a manner consistent with the capabilities they offer.

2.1.3 Ethics considerations of data collection

The rapid advancement of machine learning in medicine has prompted new questions about the **legal framework and ethics of data collection**. The **legal framework** varies by country. In the United States, the Health and Human Services (HHS) Privacy Rule standards [21] address the use and disclosure of individuals' PHI, which includes information in a medical record that can be linked to a specific individual. For research, the Privacy Rule stipulates that covered entities are permitted to use and disclose PHI (1) with individual authorization or (2) without individual authorization under "limited circumstances" that must

be approved by Institutional Review Board (IRB). In the European Union, the General Data Protection Regulation (GDPR) provides the framework for data protection and includes considerations for the use of healthcare data for a purpose different from the one for which it was originally collected (secondary use) with and without explicit patient consent. Many other countries have also established guidelines or regulations on ethics considerations for the use of human subject data [22]. For example, China released Personal Information Security Specification in 2018 to promote privacy rules established in their 2017 Cyber Security Law as a national standard [23, 24]. Brazil established the Brazilian General Data Protection Law (LGPD) in 2020; while it is broadly aligned with the EU GDPR, some notable differences exist [25]. Independent of legal considerations, several authors have recently argued for an ethical framework in which the secondary use of clinical data without explicit patient consent is ethically justifiable, as long as mechanisms are in place to ensure that ethical standards are strictly followed [26]. Additional issues related to **ethics of data collection** for machine learning systems in medical imaging include: (1) de-identification of PHI in medical images and other supporting data, and (2) impact of data collection on algorithm fairness [27].

2.1.3.1 *De-identification*

De-identification refers to removal or encoding of identifiers from patient health information collected for research purposes. In radiological imaging, many of these identifiers are present in the DICOM header contained within each image file when the image is generated for patient care purposes, and several toolkits offer a number of different strategies for de-identification of DICOM headers. For example, the Radiologic Society of North America's Clinical Trials Processor is a tool that is recommended for de-identifying DICOM headers when optimal security is required, due to its high level of customization [28]. De-identification of DICOM headers, however, may be insufficient for some radiological datasets, because there may exist potential sources of PHI other than those within the DICOM header [29]: actual pixels within the image ("burned-in" data) might contain PHI, especially in ultrasound images and radiographs; objects worn by a patient that contain personal information (such as a bracelet) may appear in medical images; and data in head-and-neck CT images may allow facial reconstruction that could identify the patient. For these reasons, it is advisable to visually inspect images and use additional tools for optimal security, especially if the images are to be publicly shared.

2.1.3.2 *Diversity and Inclusion*

A potentially significant, yet subtle, consequence of improper data collection might be an algorithm that performs poorly for certain subgroups or subpopulations with the targeted disease or condition as a result of under-representation of those subgroups in the training set [30, 31]. In radiology applications, it is important to be vigilant so that training/validation dataset selection incorporates safeguards to minimize underlying distortions for under-represented and/or vulnerable populations and so that already-existing health-care inequities are not perpetuated or exacerbated [27, 32-34].

2.1.4 Quality considerations

Image quality may have a strong impact on the reported performance of CAD-AI systems. Fortunately, many imaging centers have an image QA program already in place, and imaging exams are typically repeated if the image quality is substandard. Nevertheless, it is still good practice to ensure that a QA program is being followed at image collection sites and

to visually inspect key images to ensure image quality is acceptable before entering a case into a database for CAD-AI training, if feasible.

An additional consideration is whether the images were acquired with equipment that is still technically relevant and in accordance with appropriate image acquisition protocols. This ensures that a CAD-AI system trained or tested with the dataset is capable of answering clinically relevant questions. With rapid advances in image acquisition hardware and software, a collected dataset can quickly become obsolete. To create an enduring image dataset, data collection and management should be considered a continuous process rather than a one-shot effort.

Consideration of data curation is essential to the integrity of an image dataset. The dataset should be inspected (either visually or by automated analysis) to ensure that it contains only images from the relevant anatomic site and image modality. It is important to be aware of the differences in image acquisition parameters, imaging time points, selected series from CT scans, contrast enhancement status, and contrast administration timing. A more subtle point for data curation involves awareness of the potential bias that may be introduced if "positive" cases, for example, come from one site or scanner while all "negative" cases come from a different site or scanner, a situation that should be avoided. If developing a multi-institutional dataset, curation should be performed at the institutional level, where local clinical information is more easily accessible and verifiable, before depositing to the dataset, if possible.

2.2 Data Augmentation

Data augmentation is a collection of task-dependent techniques used to create alterations of the training data or to create synthetic data to increase the training set size aiming to improve the generalization that may be achieved by a trained CAD-AI algorithm [35]. Data augmentation has become an essential part of the training process for CAD-AI algorithms due to the recent use of deep neural networks that have millions of parameters and thus require a large number of training iterations for adequate training. To create variations of existing images contained within the training set, early successful deep learning applications for image classification used parameterized transformations that included affine transformations such as image rotation, flipping, scaling, and jittering [36]. Non-rigid transformations such as deformable transformations were later used for data augmentation.

Data augmentation based on the recently developed technique of generative adversarial networks [37] has attracted strong interest. Generative adversarial neural networks have the ability to learn the underlying data distribution and to generate synthetic images mimicking the actual ones that may fill the gaps in feature distributions [38]. Other approaches to data augmentation include obtaining images from physical phantoms or generating synthetic data from physics modeling [39]. Physical and virtual phantoms have been used in medical imaging for development of new imaging techniques, improvement of existing imaging modalities, and the conduct of virtual clinical trials; images generated from these approaches represent a natural extension for data augmentation.

Data augmentation techniques that create alterations of the training data should not modify the image appearance in a manner that makes the underlying biological or tissue properties implausible. In addition, it should be recognized that these techniques can only generate slight variations to the structural properties of existing samples in the training set; they cannot create new patterns or independent information that do not exist in the original training set. Although data augmentation may help the machine learning algorithm better interpolate among existing samples, it cannot fundamentally compensate for the inadequacies of a small clinical training set. The use of synthetic data (in silico and phantom) may prove useful for creating large training sets if the real-world variabilities of the clinical task, and the

human subjects, and the imaging system can be realistically modeled. It remains to be shown that these synthetic data can sufficiently simulate the physiological or biological properties of real patients required for developing decision support tools for many clinical tasks.

2.3 Data Harmonization

Data may include images obtained at different sites, acquired with different equipment and image-acquisition parameters, and reconstructed and/or post-processed using different algorithms. These differences may result in systematic variations across images. **Data harmonization** aims to reduce these variations retrospectively after acquisition while preserving the biological variability captured in the images [40]. Technically, DL-based methods are capable of handling variations in image appearance provided the training dataset includes example cases capturing all those variations and each in sufficient number to provide adequate learning; however, the demands of such inclusion on dataset collection and subsequent training could become prohibitively resource intensive. Moreover, deep learning methods can learn which site an image came from (for multi-institutional datasets) or which vendor's equipment was used for image acquisition, so utmost care should be taken to minimize bias in the training data [11]. For example, if all mammograms with breast cancer were acquired on a mammography unit from vendor A and all mammograms with benign lesions were acquired on a mammography unit from vendor B, a deep learning method is apt to learn to distinguish images from vendor A from those from vendor B rather than to distinguish the salient imaging features between breast cancers and benign lesions.

In practice, data harmonization has become the key to enhancing accuracy and robustness of CAD-AI systems [36, 41]. Researchers should be aware of the heterogeneity of image appearance and quality (and record, for example, differences in image acquisition parameters) during the data collection stage and incorporate data harmonization methods, when appropriate, to aid models in accommodating data heterogeneity [42, 43]. Harmonization methods can be applied in the image domain or feature-space domain [44]. Image-domain harmonization methods include post-processing of image data [45] and style transfer [46], and feature-domain harmonization methods include basic statistical normalization techniques [47] and advanced statistical techniques such as ComBat [48, 49]. The Quantitative Imaging Biomarkers Alliance (QIBA) and the Quantitative Imaging Network (QIN) have also devoted efforts to the harmonization of medical imaging data and tools [50, 51]. It is important to recognize that although data harmonization aims to reduce the systematic variations due to image acquisition, reconstruction, and post-processing or due to different protocols among data collection sites, it does not address the issue of systematic variations among patient sub-populations (see sections 2.1.3.2 and 4.2.2.3).

2.4 Take Home Message on Data

In summary, proper data collection methods are of critical importance to successful training, validation, and implementation of CAD-AI algorithms. Improper collection and manipulation of data (such as improper data augmentation) can lead to an overestimation of performance or lack of generalizability.

3 Reference Standards

The development of machine learning-based decision support tools requires truth or labeling of the cases for training, validation, and independent testing. The resulting reference standard needed for the evaluation of an algorithm's (or human's) performance depends on the task at hand. It is important to note that the notion of "truth" (or "ground truth" or "gold

standard") has been replaced by the concept of "**reference standard**," as very few, if any, real-world tests yield the absoluteness implied by "truth" or "gold standard." In many respects, the clinical utility of an algorithm greatly depends on the quality of the reference standard used in its training and evaluation. It is challenging but crucial for investigators to (1) select the most appropriate approach to obtain a task-specific reference standard, (2) gather complete and reliable data for that reference standard, and (3) assess any biases that may be introduced when training their algorithm with a reference standard that contains inherent variability.

This section covers considerations for generation of reference standards including objective vs. subjective reference standards, annotation granularity**,** methods for acquiring annotations, definition of true positives. The use of the reference standard in training and model development (section 4) and in performance evaluation (section 5) of a CAD-AI algorithm are closely related.

3.1 Objective vs. Subjective Reference Standards

The most straightforward reference standard uses the collected image data itself, with one or more domain experts providing diagnostic assessments or annotations at the image or patient level. **Reference standards based on physicians' opinion, however, are subjective**, and several studies have shown that CAD-AI system performance may vary substantially when assessed against different reference standards provided by radiologists [52-57]. Subjective reference standards are considered more reliable if they are based on consensus of multiple experts; however, it is difficult to estimate the number of experts needed. Ideally more than two experts should participate in order to identify outliers. It can be expected that the preferred number of experts depends on the task for which the reference standards will be used, the difficulty of that task, and the expected variability of the generated reference standard. In practice, obtaining a reference standard from experts is a very resource-intensive task, and usually only limited expert readings are possible, especially for large datasets.

Further reliability for reference standards may be achieved with information from other independent sources [58, 59], which also may be consensus based, such as radiologist's review of images from another modality [60] or imaging follow-up for 2 years or longer [61].

Despite the prevalence of subjective approaches that use expert opinion, more **objective reference standards** are frequently desirable. For example, for lesion detection and pathologic classification, more definitive diagnostic tests and pathologic assessment of biopsied or excised lesions [62], although imperfect, should be used. For clinical decision support, such as treatment response assessment or patient prognosis, a more objective reference standard is patient survival. While the date of patient death is definitive, procuring this information as a reference standard becomes complicated by the need to track patients over potentially extended periods of time, during which they might become lost to follow-up; patient death could also result from circumstances other than the disease being evaluated. Shorter-term reference standards such as time-to-progression also may be used as an alternative in many studies.

3.2 Annotation Granularity

The level of required **annotation granularity**, or detail, depends on the task. For example, a more object-specific annotation such as manual expert delineation may be needed for lesion/organ detection or segmentation. For diagnosis of systemic disease or patient prognosis, patient-level assessment or patient survival may be appropriate. Image-based

reference standards of varying levels of granularity are the most commonly used ones for current medical imaging-based machine learning tasks.

3.2.1 Entire image

The coarsest level of granularity is **annotation of the entire image**, through which a class label is assigned to each image. As an example, the DREAM Challenge [63] for digital mammography diagnosis only had available breast-level labeling regarding the presence of breast cancer; however, training with such global labels that do not locate the actual lesions is sub-optimal in guiding deep networks to learn the relevant features of those lesions that are responsible for the patient-level diagnosis[1]. The top-scoring teams in the DREAM Challenge all used additional datasets with lesion location labeling to supplement the training of their systems. Another study showed that without specific lesion locations, the system could learn non-medical features that were included in the images (such as metal labels and markers), thus impeding the generalizability of the algorithm [11]. A more recent study [64] showed that the performance of an AI system for screening mammography on unseen cases varied from modest to outstanding depending on the dataset and reference standards used for evaluation.

3.2.2 Region-based

A finer level of granularity is annotation of specific lesions or organs through expert manual marking of a bounding box or a region center point. If the purpose is to detect cancers, for example, the CAD-AI system has to characterize the level of suspicion of a potential target structure and mark it as a cancer if it satisfies a certain threshold suspicion level. The scoring of system performance, then, requires not only the location of the lesion as reference standard but also the established malignancy status.

3.2.3 Pixel-based

An even finer level of granularity is **pixel-based annotation** in which the reference standard is an expert manual delineation, or outline, of the lesion or organ of interest and each image pixel can be labeled as either belonging to the region of interest or not. These detailed annotations are important for evaluating performance when the task is organ or lesion segmentation, and they can also be important for applications such as lesion characterization or treatment response assessment, in which the lesion extent and radiomic features are extracted from the segmented lesion. Pixel-based reference standards are more detailed than region-based ones but come at the cost of a more time-consuming annotation process and larger inter-reader variability [65].

3.3 Methods for Acquiring Annotations

3.3.1 Expert labels

When clinical or pathologic information is not available, it is common (for certain CAD-AI tasks such as lesion detection or segmentation) to create **a subjective reference standard from human domain experts**, who label the images or mark individual pixels, depending on the level of annotation granularity required. Outlining the boundaries of lesions or organs has the disadvantage of requiring potentially extensive time and effort, especially for manual segmentations in 3D. The judgment of lesion boundaries or the presence of a lesion contains

[1] Recent "weak learning" and "attention" mechanisms may provide solutions for this (see Section 4.2)

intra- and inter-observer variability, even for experienced radiologists [65, 66], so that multiple experts may be required to produce a reliable reference standard.

3.3.2 Electronic health record

For patient-level assessments, the **electronic health records (EHR)** of subjects can be parsed by humans or natural language processing algorithms for reference standards involving, for example, the presence or absence of disease. Reference standards obtained from EHR data may contain annotations made during clinical practice, such as bounding boxes or Response Evaluation Criteria in Solid Tumors (RECIST) measurements [67]. If performed manually, a reference standard obtained from the EHR is time consuming and may not be practical for collecting large datasets; if performed automatically, the labels may contain a lot of noise and be prone to error, especially for complicated cases [68]. Natural language processing for parsing EHR data is an area of active research. It should be noted, however, that clinical radiology reports are not recommended as a reference standard for CAD-AI development, because "clinical reports often have nuanced conclusions and are generated for patient care and *not* for research purposes" [69].

3.3.3 Crowd sourcing

The key concept of **crowd sourcing** is to switch the time commitment and required effort for a given task from domain experts to many, potentially less-experienced, users. Crowd sourcing is a form of subjective consensus reference standard that has been applied to image annotation, image segmentation, and object delineation tasks [70]. It has been shown, in certain settings, that the quality of annotations from experts and those from novices becomes equivalent with an increased number of novices [71, 72]. Nevertheless, the use of crowd sourcing as a reference standard for machine-learning applications in medical imaging must be further investigated before it can be recommended for general use.

3.3.4 Phantoms

In medical imaging, **phantoms** are man-made objects with known structure and composition. Images acquired of these phantoms support *a priori* image annotations across a range of granularity levels. However, the number of physical phantoms usually is limited, and, therefore, only a few annotated images can be obtained from this method. Recently, digital phantoms that mimic properties of physical objects *in silico* have become available [73] and have been used in virtual clinical trials [73, 74] as well as for training ML models [39]. An advantage of using *in silico* models is that the lesion location and properties are known by design so that human annotation is not required; however, image data obtained from phantoms (physical or digital) likely do not reflect the actual biological or pathological characteristics that may be captured on patient images. Phantom images may be useful for data augmentation during training, for identifying and correcting biases regarding differences in imaging systems and protocols, and for test-retest evaluations. Whether an algorithm trained with phantoms is applicable to real-world images requires rigorous validation [39]. Similar caution must be applied to the use of synthetic images generated by digital methods such as full *in silico* modeling of the imaging chain or use of generative adversarial networks.

3.3.5 Weak/noisy labels

Weak or noisy labels can be defined as incomplete or imperfect reference standard annotations. Compared with a small dataset with "strong" or "clean" labels, a large dataset with "weak" or "noisy" labels used for algorithm training may achieve comparable performance [72]. The generalizability of the trained algorithm, however, will deteriorate as the proportion of noisy labels in the training set increases [75]. Others have demonstrated the

potential of using weak or noisy labels [76] but additional research is needed. Strong labels specifically for the independent test set are essential to reliably assess the performance of the trained decision support tool. Under certain circumstances, the STAPLE algorithm ("Simultaneous Determination of a Reference Standard and Performance Level Estimation") delivers not only the optimal reference standard estimation but also a quality ranking of the competing observers/algorithms [77].

3.4 Definition of True Positives

Reference standards are designed for use in evaluating the output of a CAD-AI system. The definition of a **true positive** relative to the reference standard is very important. Different methods for determining a true positive will result in different performance of the same CAD-AI system. Which method is appropriate or feasible depends on the task and the available reference standard. Using detection tasks as a specific example, a number of methods have been used to determine whether the lesion is correctly detected, including the distance between the centroids of the detected object and the reference, the overlap percentage between the two (which is further affected by the level of detail in marking the reference, e.g. bounding box vs. outline) [78], and whether the centroid of the detected object falls within the reference lesion region; detected objects that are not determined to be true positives through the selected metric are counted as false positives. It has been shown that scoring is strongly affected by the detection criterion [79]. More detail on performance evaluation can be found in section 5.

3.5 Take Home Message on Reference Standards

The required type and granularity of the reference standard depends on the task at hand. An objective reference standard is preferred; however, when a subjective reference standard cannot be avoided, independent assessments of multiple domain experts should be obtained and their variabilities should be evaluated.

4 Model Development

In addition to the availability of properly collected data and labels, the selection of data sampling and machine learning strategies will affect the robustness of the developed model. This section covers the topics of data sampling methods, levels of learning supervision, and new training strategies, including transfer learning, multi-task learning, domain adaptation, federated learning, and continuous-learning. A recent review on some of these technologies and their applications can be found in the literature [80].

4.1 Data Sampling Strategies

Data sampling is important for efficient use of data and for reducing the risk of overfitting in model development. The most established resampling techniques for the training and testing of models will be discussed. The dataset ideally should be split into three non-overlapping partitions: **training**, **validation**, and **test** sets. One of the partitions should be used for training of the model. To guide the optimization (or tuning) of model parameters during training of a model, it is desirable to obtain a meaningful estimate of the performance of the model being trained on a partition of the dataset that is often referred to as a "validation set;" the use of the validation set is thus a part of the training process. This is not to be confused with the use of the term "validation" as the process of evaluating the generalizability of a developed model on unseen data after training is completed and the

model is "frozen," which should be established by **testing on a completely independent dataset** from the ones used during the training or optimization of the model. To avoid overfitting the model, performance testing ideally should be conducted only once on any given **test set**; the performance on that test set should then not be used to inform model improvements or modifications for subsequent testing on the same test set [5, 14, 81]. Due to potential confusion surrounding the term "validation" for reporting the performance of a trained model, developers need to clearly define whether the test set used for the evaluation has been kept independent from the training process. There are several established resampling techniques for organizing the training and evaluation of a model, especially with limited datasets. It should be noted that such techniques are generally based on the assumption that the available data are representative of the underlying target population and similarly distributed within the training, validation, and test datasets.

A **holdout method** is the most basic evaluation/training paradigm. In this approach, a model is trained and optimized by use of training and validation datasets, after which it is evaluated once with an independent test dataset that is sequestered during training. When the available datasets are small, a **k-fold cross-validation** method, which maximizes the use of the available data, can provide a more reliable evaluation of model performance than the holdout methods under this condition [82, 83] if the test partition in each fold is held-out as an independent test set and is not used repeatedly for guiding model optimization. For such techniques, stratified sampling of cases (Section 2.1) can better accommodate imbalanced datasets than random sampling. **Bootstrapping** is another popular and well-established resampling method that can be used to construct sampling distributions for model training and evaluation purposes [84-86].

Although the actual generalization performance of the final model should be evaluated only once by external testing with a previously unseen independent test set, in practice, it is psychologically difficult for researchers not to go back and improve their model if the observed test performance is poor. Such multiple testing and reuse of the same test data are likely to introduce overfitting problems regardless of the evaluation/learning paradigm [81, 87].

4.2 Machine Learning Strategies

A machine learning paradigm refers to a strategy based on which a model is trained. There are numerous learning paradigms in CAD-AI, many of which overlap [88-90]. One approach for categorizing learning paradigms focuses on the level of interaction required by the user, such as supervised, semi-supervised or unsupervised learning. A different approach considers the learning paradigm from the perspective of model development, such as transfer learning, multi-task learning and federated learning.

4.2.1 Levels of learning supervision

Supervised learning (with different levels of supervision) is the most common approach to learning, where a model is trained to map input data to output data based on examples of the input-output pairs. To reduce the cost and barriers related to data collection and annotation, however, several studies are actively exploring training algorithms that can leverage unlabeled or weakly labeled data during training (see also Section 3.3.5). Such paradigms may provide a more cost-effective and scalable approach to CAD-AI development.

4.2.1.1 S*upervised learning*

In **supervised learning**, a model is trained to map input data to output data based on explicit examples of the desired input-output pairs, as provided by the user. However, the

collection of such annotations tends to be costly and time-consuming, and the annotation effort may need to be repeated as the imaging technology evolves and new datasets are introduced. Moreover, as noted in previous sections, annotations can be subjective, the annotation process may be prone to error, and, for some tasks, there is no single correct annotation.

4.2.1.2 *Semi-supervised learning*

Semi-supervised learning algorithms exploit a combination of labeled and unlabeled data. In this case, the model is given some guidance about the desired outcome, but the annotations do not need to be as detailed or extensive as those used with supervised learning. For instance, feature extraction can be initialized through an unsupervised or self-supervised technique and then fine-tuned to the final task using a small set of labeled data. Using some form of semi-supervised learning may reduce the costs of labeling relative to supervised learning.

4.2.1.3 *Self-supervised learning*

Self-supervised learning can exploit large unlabeled datasets for feature representation and has a regularizing effect on the learned features. Autoencoder models are a common approach to self-supervised learning [37] and are used for feature extraction; however, there is no guarantee that the features learned in a self-supervised fashion have diagnostic value. It should be noted that autoencoder models, such as U-Net, can also be used in a supervised mode for image segmentation tasks. Other popular approaches to self-supervised learning include *contrastive learning* [91-93] and *pretext* [91] *or surrogate supervision* [94]. In these techniques, when a large unlabeled dataset in the same domain as a small labeled dataset is available for a given task, the unlabeled data can be assigned artificial labels and then used to pre-train a deep learning model; transfer learning for the target task is then performed with the small labeled dataset. It has been shown that deep models pre-trained with self-supervised learning techniques can outperform the same models trained with random initialization [95] or transfer learning from an unrelated domain [94, 96]. These findings demonstrate the potential of large datasets to improve model development in medical imaging tasks even if a large portion of the cases is unlabeled.

4.2.1.4 *Unsupervised learning*

Unsupervised learning refers to a class of algorithms that can autonomously learn from data without reference to any labels or any instruction from the user. Common approaches to unsupervised learning are the clustering methods. Unsupervised learning has shown promise in medical imaging applications but depends on the adequacy of the resulting automatic clustering. In addition, unsupervised learning requires a much larger training set for the algorithm to achieve similar performance compared with training with reference standard [97], and data collection in medical imaging is costly.

It should be noted that CAD-AI algorithms can include both supervised and unsupervised elements.

4.2.1.5 *Multiple-instance learning*

The **multiple-instance learning** approach is an effective paradigm when labels are not available at the desired granularity [98]. The machine learning model receives a set of labeled "bags," each containing many (unlabeled and some labeled) instances. In the simplest case of binary classification, a bag is labeled positive if it contains at least one positive instance.

4.2.2 Transfer learning, multi-task learning, and domain adaptation

The ability to discover by **representation learning** a wide range of object characteristics is a distinctive advantage of deep learning over traditional machine learning models that rely on hand-engineered features [99]. In deep convolutional neural networks (DCNNs), feature extraction is obtained through a series of cascaded convolutional layers, forming a hierarchy in which shallow layers extract generic features and deeper layers extract increasingly object-specific features [100]. Large-scale datasets, however, are needed to learn high-quality features, thus making deep learning an effective, but data and computation hungry, paradigm. Such data requirements can be lessened by transferring or sharing features across different tasks and domains.

4.2.2.1 *Transfer learning*

Transfer learning in DCNNs is commonly implemented by training a network on one task and then "transferring" the parameters (or weights) from the trained model to initialize the network for a new task, rather than randomly initializing it (also known as "training from scratch"). Transfer learning was the early enabler for the use of deep networks in the medical imaging domain. Networks pre-trained on ImageNet, which comprises millions of non-medical images effectively labeled by crowd sourcing, are commonly used to initialize DCNNs for medical image classification, showing improved classification performance and faster convergence compared with random initialization [98, 101-105]. Transfer learning, however, imposes limitations on the DCNN, since ImageNet is composed of low-resolution 2D RGB color images, whereas many medical imaging modalities are higher-resolution 3D, 4D, or multi-parametric. One of the most common techniques for bridging the two domains involves a 2.5D approach [106], in which a 3D (or higher-dimensional) image around a voxel is subsampled into multiple 2D images, which are then fed into the input channels of a 2D DCNN [102] or an ensemble of 2D DCNNs [107].

For some tasks, such as segmentation, 3D convolutional filters may offer substantial advantages over 2D CNNs; in such cases, training from scratch or transfer learning from another medical imaging modality may be performed. Researchers have begun to explore medical imaging-based pre-training of DCNNs, and results indicate that an additional stage of pre-training with data from a similar domain can increase performance and robustness of a network [108, 109]. The transfer of prior knowledge can occur between modalities (e.g., CT to MRI), between organs/pathologies (e.g., liver to kidney), between tasks (e.g. classification to segmentation), or some combination thereof [110].

4.2.2.2 *Multi-task learning*

Multi-task learning is a special type of transfer learning in which a DCNN is trained to jointly learn interrelated tasks, as opposed to addressing each task sequentially [111]. This technique has demonstrated enhanced performance compared with single-task learning [110, 112].

4.2.2.3 *Domain adaptation*

Most algorithm training methods assume that the test data is drawn from the same distribution as the training data; however, this assumption is often not fulfilled in practice due to data scarcity and data mismatch, and thus a trained model may fail to generalize to real-world clinical data [113, 114]. The most important sources of **data shift** (i.e., deviations between the distributions of the test set data and the training set data) in medical imaging are acquisition shift and population shift (Table 1) [11].

Data shift can be addressed, at least partially, through data harmonization and standardization, as discussed in Section 2.3. Recently, researchers in the medical imaging space have begun to explore domain adaptation techniques to make deep learning models more tolerant of domain shift [115]. The most common approaches to domain adaptation are feature based and attempt to modify the feature distributions to align the target (i.e., test set) and source (i.e., training set) domains. Other approaches seek to learn domain-invariant representations [116] or use generative models to synthesize realistic samples in target domains where labeled data are scarce [117-120] [38].

4.2.3 Federated learning

Federated learning is a distributed machine learning approach that enables collaborative training on decentralized datasets [121-124]. Each site trains the model locally with its own dataset and then only the trained model parameters are shared, thus producing a global model benefiting from access to a large corpus of data without requiring data sharing and without posing risks to patient privacy. There are, however, several open-ended questions with regard to federated learning that are relevant to medical imaging [125, 126]. In particular, there is no formalized training protocol yet to guarantee that the performance of a model trained with federated learning is comparable to that of a centralized trained model with access to all the data [127]. Also unknown is (1) the extent to which local model overfitting negatively impacts the global model, and (2) the tradeoff between access to more data through a federated process versus traditional learning with a fully controlled dataset.

4.2.4 "Continuous learning" systems

Continuous or "life-long" learning emulates the human ability to continuously learn and adapt as new data are presented [128, 129]. Theoretically, continuously learning AI systems can accelerate model optimization and continuously improve their performance by taking advantage of new data presented during clinical use. In practice, adaptive training of shallow and deep neural networks using incrementally available data generally results in rapid overriding of their weights, a phenomenon known as "interference" or "catastrophic forgetting" [130, 131]. It is not generally clear under what conditions and for what metrics adaptive AI produces a continuously improving (or at least stable) algorithm and avoids major pitfalls. Many questions related to post-marketing management of adaptive AI devices remain open, such as frequency of adaptation (e.g., continuously or in regular intervals, batch mode), how to monitor performance changes after adaptation, and when and how to intervene if performance decline is suspected.

4.3 Take Home Message on Model Development

Training approaches, especially for deep learning algorithms, are continuously improving with the goal of achieving robust, effective, and privacy-preserving CAD-AI models. An independent test set representative of the intended use that was not employed to guide model optimization in any learning paradigm is of critical importance. Robust training methods, although important for all CAD-AI systems, are especially important for systems that may operate in clinical practice with minimal or no human supervision.

5 Performance Assessment

Proper performance assessment is important in various stages of CAD-AI model development. Performance assessment involves (1) factors such as intended use, performance metrics, statistical significance, sample size, and reproducibility and (2) purposes such as

standalone or clinical reader performance assessment. Rigorous performance assessment can provide a reliable estimate of model performance at a particular stage of development to guide further improvement or to inform the user of realistic performance that can be expected from the model. This section discusses methods and considerations for conducting performance assessments.

5.1 Performance Assessment Metrics

In CAD-AI applications, the most widely accepted performance assessment methodologies include receiver operating characteristic (ROC) analysis [132], its various derivatives such as free-response ROC (FROC) analysis [133], and precision-recall analysis. In detection and classification tasks, the most common metrics include area under the ROC curve, sensitivity (or recall), specificity, balanced accuracy (mean of the sensitivity and specificity), Youden index, and the prevalence-dependent factors positive predictive value (or precision), negative predictive value, and F1 score [5, 134, 135]. Various other methodologies and metrics have been established for specific applications, such as the Dice coefficient, Jaccard index, and Hausdorff distance for image segmentation; mean squared error and coefficient of determination for regression; concordance index [136, 137] for evaluating prediction performance; the log-rank test [138] for comparing Kaplan-Meier survival curves in survival analysis; and categorical agreement of response classification by, for example, the RECIST guidelines [139, 140]. The use of multiple performance approaches is generally appropriate to provide a more complete assessment.

It is crucial to include error estimates, such as standard deviations or 95% confidence intervals, when reporting results. Error estimates describe the uncertainty/variability of the reported values for the performance metrics and help provide insight into the sufficiency of the training sample size, the soundness of the training/testing approach, and generalizability.

5.2 Statistical Significance

Statistical significance is used to quantify the likelihood that an observed result is explainable due to chance alone [141]. **Statistical power** is a closely related topic that quantifies how likely a study is to distinguish an actual effect from one of chance. Whereas statistical significance of results is assessed *after* study completion, statistical power calculations are an important part of study design and performed *beforehand* to estimate the required sample size based on the expected size of the effect, variability in the response variable, and disease prevalence [142]. Failure to achieve a statistically significant result cannot be interpreted as a true lack of difference especially when the study is statistically underpowered. It is important to note that statistical significance does not necessarily imply that the result is clinically meaningful [143, 144] unless the study is specifically powered to address this issue. Moreover, when multiple statistical hypotheses are tested using the same dataset, the chance of observing a rare event increases, thereby increasing the likelihood of incorrectly concluding that a real effect has been observed when the observation, in fact, was due to chance alone; methods for adjusting for the effect of multiple hypothesis testing have been developed [145]. Statistical tests generally make a set of assumptions about the distribution of the data to which they are applied (e.g., normality or linearity), and it is important to verify these assumptions are met before using any specific statistical test.

5.3 Intended Use

The **intended use** for which a CAD-AI system is designed must match the clinical environment in which it is deployed. The intended use is determined by the patient population, the image acquisition device, the stage of diagnostic intervention, and the diagnostic category. First, the patient population represented by the data used to develop the

algorithm should match the intended population. Second, a range of image acquisition devices are in clinical use, and CAD-AI must be developed and tested on data from multiple vendors. Third, the intended use depends on the patient care stage that requires the diagnostic intervention. Finally, the diagnostic category of the data should match the clinical task, for example, screening, detection, staging, treatment assessment, or follow-up.

CAD-AI systems for aiding in clinical decision making generally may be implemented according to four different paradigms: **second read, concurrent read, triage, and rule-out**. CAD-AI applications such as detection and diagnosis as well as staging, treatment response assessment, prognosis, or recurrence prediction (Figure 1) should be matched with the most appropriate paradigm. The selected performance assessment method should be reflective of the use paradigm (Table 2). Frequently, the setting may affect the operating point of the CAD-AI tool, e.g., the relative importance of sensitivity vs. specificity. In addition, CAD-AI systems designed for different intended uses may have different performance requirements; for example, CAD-AI systems designed for disease detection in a concurrent-read paradigm generally should have higher sensitivity and specificity than those used in a second-read paradigm due to potentially increased reader reliance on the computer output in the former setting. CAD-AI devices that operate at performance levels that rival those of human experts [146-148] could potentially be the basis for future autonomous AI devices that bypass human interpretation in selected cases or for selected tasks. An example of such applications is **rule-out devices**, a class of devices designed to identify and remove negative cases without clinician review. Although some authors have considered rule-out as a subset of the triage paradigm, the clinical implementation of each requires a unique set of strategies and performance assessment considerations due to different levels of risk associated with each approach.

5.4 Standalone Performance Assessment

The evaluation of a CAD-AI algorithm includes both benchmarking algorithm performance and assessing the added value to the end user provided by the algorithm in improving clinical decision making [5]. **Standalone performance assessments** are employed during development to allow for modifications to be quickly compared to previous models. For benchmarking, overall performance is based on an independent dataset representative of the clinical population acquired using the expected range of image acquisition technologies and protocols for the intended use.

5.5 Clinical Reader Performance Assessment

A **clinical reader performance assessment** is used to estimate the clinical impact of a CAD-AI algorithm [153, 154]. A common approach for assessing clinical performance is through a controlled reader study (either retrospective or prospective), directly comparing the performance of a human reader without and with output from the CAD-AI system [155, 156]. A disadvantage of this approach is that the estimated performances are unlikely to match those in the true clinical setting because of differences in the cases, physicians, and reading process. It is important to realize that both the population of patients undergoing the examination (cases) and the population of physicians interpreting the data (readers) are sources of substantial variability in clinical reader studies [157]. Specialized statistical and methodological tools are needed for these analyses [158]. Well-designed clinical reader studies can be used to gain Food and Drug Administration approval (or approval by a similar organization outside of the United States) for clinical use of a CAD-AI system and are often a precursor to direct assessment of diagnostic performance in clinical practice (Section 6.4.3).

5.6 Sample Size

Assessing performance dependency on the training **sample size** in medical imaging is important to achieve a viable clinical translation. As previously discussed (Section 4.1), small training sample sizes may lead to overfitting, or overtraining, of CAD-AI algorithms. In general, the performance of CAD-AI systems depends on the training sample size, disease prevalence, the number of features and their statistical distribution, the choice of the CAD-AI model, and the scoring metric [82, 85, 159, 160]. For the deep learning techniques, the training sample size is even more critically important since millions of parameters need to be determined. Even when deep learning models are trained with transfer learning (Section 4), the training sample size is still a major factor that affects performance and generalizability. The variability in the algorithm performance from repeated experiments at different sample sizes can be used to assess overfitting and generalization error [75, 108].

5.7 Reproducibility

It is important to clearly specify the conditions under which the results of a CAD-AI system are reproducible. Recent studies have distinguished among different types of reproducibility [161-163]. Three types of reproducibility have been defined, the first two of which are relevant for model validation and successful clinical deployment of CAD-AI systems.

Technical reproducibility refers to the ability to precisely replicate reported results (usually in a publication) based on a complete description of the method and release of the corresponding code and dataset.

Statistical reproducibility refers to a result being valid (within a specified standard deviation or confidence interval) when different variations of the training conditions are applied. Variations in training conditions will result, for example, from different random seeds, from different partitions of the training set, or from different strategies to divide the dataset into training and test subsets. Statistical reproducibility in model performance will also depend on the test set. If different test sets are sampled from the same population, the DCNN output will be different for the different test sets due to statistical variation of the test sets. If the test is repeated multiple times, and each time a different test set is randomly drawn from the population or by bootstrapping, the test performances can be considered samples from the same statistical distribution, from which the mean performance and standard deviation can be estimated.

Inferential reproducibility refers to the ability to reach qualitatively similar conclusions from an independent replication of a study under conditions that match the conceptual description of the original study.

5.8 Take Home Message on Performance Evaluation

The most appropriate performance metric(s) will depend on the task and the reference standard. Often multiple performance metrics are appropriate, and use of multiple metrics is frequently desirable. Power calculations should be an integral part of study design, and performance analysis should include error estimates, assessment of statistical significance, and preferably assessment of reproducibility.

6 Translation to Clinic

The ultimate goal of developing CAD-AI system is to assist physicians in the health care process. For clinical acceptance of a CAD-AI tool, many practical factors must be considered, such as generalizability to the clinical environment, efficiency of use in a clinical

workflow, explainability of the output, and assurance of performance consistency over time. This section will discuss topics related to the translation of CAD-AI tools to the clinic, including human-machine interface, user training, acceptance testing, and prospective surveillance.

6.1 Human-Machine Interface

One of the most important issues of introducing CAD-AI to clinical use is the presentation of its output to the physician. The **human-machine interface** is a critical component that can impact the usefulness and the acceptability of a CAD-AI tool for clinical use. The interface design will depend on the intended use (e.g., disease detection, triaging, treatment response assessment); the amount, type, and complexity of information to be displayed (e.g. markers, parametric maps, likelihood scores); the reader paradigm; and the level of interactivity (e.g., when and how the physician can enable, disable, or query the CAD output). Regardless of the task, some common requirements may include user friendliness, workflow efficiency, and the interpretability of the CAD-AI output or recommendations.

The black-box nature of current CAD-AI tools is one of the roadblocks to translation of CAD-AI into clinical use. Providing uncertainty estimates of the output could allow a better understanding of the black box model and improve the safety of deep learning systems [164-168]. For physicians to have confidence in a recommendation by a CAD-AI tool, it is helpful for them to understand the reasons behind the prediction or decision. The explanation has to be consistent with medical knowledge or supported by clinical evidence. **Explainable AI (XAI)** is an emerging machine learning area [169] that seeks to design interpretable AI models or, more commonly, provide post-hoc explanations for trained AI models; the most common approaches at present include generating visual heatmaps, providing examples of similar lesions or cases, and providing semantic textual explanations or cues [170]. A visual saliency map or a color heatmap of the image [171], which captures the relative contribution to the DCNN output score from various image locations, can be generated using a gradient-based, perturbation-based, or class activation map-based (CAM) method [172-176]. The local interpretable model-agnostic explanations method (LIME) [177] similarly identifies the extent to which regions or pixels influence the particular prediction. The visualization provides some evidence of the correlation of the deep features and the output score to the input data; however, visualization maps (which are generally difficult for humans to interpret) are far from a complete explanation of why and how the features are connected and weighted to identify the target lesion [169, 176]. Saliency map techniques often cannot meet key requirements for utility and robustness, emphasizing the need for additional validation before clinical use [176]. For clinical tasks more complicated than lesion detection, the CAD-AI tool may need to provide explanations or references that correlate the recommendation with the patient's medical conditions or other clinical data. Much more research and development are needed to determine physicians' preferences regarding user interface design for each type of application so that CAD-AI models can truly become intelligent decision support tools.

6.2 User Training

In translating technology to the clinic, an important step is to set expectations. Key to a **user's proper use of a CAD-AI tool** is an understanding of the intended use, including the purpose and when and how it should be used in the radiology workflow [178]. For example, if a CAD-AI tool is developed for lesion detection, the user should be informed about whether the tool is designed and validated for use in a concurrent-read or second-read paradigm. CAD-AI tools designed for different intended uses may have different performance requirements; for example, CAD-AI systems designed for disease detection in a

concurrent-read paradigm generally should have higher sensitivity and specificity than those used in a second-read paradigm due to potentially increased reader reliance on the computer output in the former setting.

A second key issue is to **acquaint the user with both the capabilities and limitations of a specific decision-support tool**. Users should have a comfortable level of trust in the CAD-AI tool but should always be aware of the performance limitations of the tool. The performance of a CAD-AI tool can be affected by patient demographics, imaging equipment, and image-acquisition protocols. Even if a CAD-AI tool has been trained by the vendor with multi-institutional data and approved for clinical use, its performance in the local population may not be the same as that specified by the vendor. An initial user-training and adjustment phase is recommended as an integral part of the deployment. During this phase, physicians should evaluate the performance of the CAD-AI tool on their patient cases by comparing with clinical outcomes to understand the characteristics of the cases for which the CAD-AI provides correct and incorrect recommendations, but they should refrain from being influenced by the CAD-AI output in their clinical decisions. This adjustment phase will provide the user with a deeper understanding of the CAD-AI performance in the local setting, and also impart to the user an appropriate level of confidence in the recommendations generated by the decision-support system, which may reduce unrealistic expectations and improper use of a CAD-AI tool. For example, misusing a tool intended to be a second opinion as a concurrent reader may lead to disappointing outcomes, user dissatisfaction, and, most importantly, potential harm to patients [179]. The length of this training period may depend on the type of CAD-AI application, the level of risk, and the observed performance and consistency of the CAD-AI tool. The resulting insights may also provide useful feedback for the CAD-AI vendor [6].

6.3 Acceptance Testing

CAD-AI software to be implemented for clinical use is considered a medical device; its performance, therefore, must meet certain standards. **Acceptance testing** is an important step prior to clinical use of any CAD-AI tool [6, 178]. Manufacturers must provide instructions for use with detailed guidance on system installation, acceptance testing, acceptance criteria at installation and subsequent upgrades, and periodic QA. The instructions for use must also include a description of the expected performance levels of the CAD-AI system along with tolerance limits and a graphic presentation of CAD-AI output layout and proper user interface configuration.

A **basic level of acceptance testing** may use pre-collected data provided by the manufacturer or phantoms for testing the operation and consistency of certain CAD-AI functions after installation and compared with the expected outcomes. **Another level of acceptance testing** should include a set of clinically representative cases collected by the individual clinical site. The deviation of the resulting performance level from the performance level claimed by the CAD-AI manufacturer must be within specified tolerance limits. For clinical sites that may not have a large set of cases readily available for acceptance testing, the clinical performance assessment may be conducted during the user training phase, which may be less quantitative but has the advantage of being most consistent with the clinical operations at that site.

6.4 Prospective Surveillance

6.4.1 Periodic quality assurance

The goal of **periodic QA** is twofold: to establish a schedule of routine QA and to assure the consistency of clinical performance over time. Routine QA should be implemented

(preferably by medical physicists in conjunction with routine QA testing of related medical imaging systems) to assess how variations in the imaging or data collection chain may affect the performance of the CAD-AI system [6, 178]. QA should also be performed whenever a CAD-AI software update occurs, which should always be announced by the software development company. The use of phantoms for this testing is recommended if the CAD-AI system is designed to be applicable to specific phantoms and its performance has been shown to be sensitive to the quality of images acquired from these phantoms. To evaluate performance consistency in routine clinical cases, clinical sites and CAD-AI manufacturers should develop tools to track performance levels of certain indices and monitor deviations (e.g., a tool to track the number of markers per image for detection tasks [6]).

The tolerance limits and corrective actions for any observed deviations should be established based on the CAD-AI application. The risk associated with any deviation will vary significantly for different diseases and tasks performed by the CAD-AI system. For example, if the system is an autonomous CAD-AI detection or decision tool for triaging or rule-out, immediate corrective actions are recommended, while tools designed only to provide second opinion or supplementary information may be less urgent. Regardless of the risk level, awareness of these deviations by the physicians is critical as they may need to adjust their level of trust on the CAD-AI recommendation when performing clinical tasks.

6.4.2 Performance monitoring for "continuous learning" systems

For continuous learning CAD-AI systems implemented in the clinic, an additional risk results from learning from non-stationary data that may lead to catastrophic forgetting and degraded performance unbeknownst to the physicians in their daily use of the system [129]; furthermore, system performance may be frequently changing, which impacts its safety profile. The manufacturer or the in-house development team must have well-defined QA procedures to validate the quality of data, including collateral information (e.g., clinical outcomes), and assess model performance after each update. Before continuous learning CAD-AI systems can be translated into the clinic, extensive work is required to develop practical and reliable QA methods that enable performance monitoring to ensure safe use.

6.4.3 Prospective evaluation of CAD-AI

Large-scale prospective performance assessment of CAD-AI systems will evaluate the impact of CAD-AI on workflow efficiency, physician performance, cost-effectiveness, and patient outcomes in the clinical setting. Prospective evaluation of CAD-AI typically falls into two categories: **randomized controlled trials (RCTs)** and **observational studies**.

RCTs are designed to control for sources of bias through randomization, blinding, and allocation concealment. RCTs are logistically difficult to organize and generally require a large patient population. A common design is the sequential study, in which the physician interprets each case first without the assistance of CAD-AI and then, after formally recording his or her findings, interprets the case again while reviewing the CAD-AI recommendation [180-186]. This sequential design, however, cannot be applied with concurrent-read or triage paradigms, as discussed in Section 5.3 (Table 2).

Well-designed **observational studies** can be highly informative and much easier to conduct than RCTs [187]. The most common design is the historical-control study, in which the performance of groups of radiologists over different periods of time is compared; the patient cohorts and radiologists involved may not be identical for the two time periods. Observational studies are commonly used when a new predictive or diagnostic CAD-AI system has been available in clinical practice for some time after regulatory approval [188-191]; however, care must be taken to account for differences such as the characteristics of the patient population and physicians' experience between the two time periods, since such

differences may bias the observed outcomes. Relevant statistical procedures such as stratification and multivariate regression modeling can be used to account for confounding factors.

The reporting of a clinical trial evaluating a CAD-AI system in the literature should allow readers to identify potential sources of bias and, ideally, reproduce the results. Factors that may bias or impact the results include the study population, data acquisition, characteristics of the CAD-AI device, human-AI interaction, user training, study end-point, reference standard, and statistical methods, all of which should be clearly identified and reported. Additionally, the SPIRIT-AI [192] and CONSORT-AI [193] extensions provide general guidelines when drafting clinical trial protocols or reports that target or include CAD-AI systems of any kind. It should be noted that the CONSORT-AI statement does not yet cover advanced learning paradigms such as continuously evolving or adaptive systems, the performance of which may change over time, and underscore the importance of a robust post-deployment surveillance plan.

6.5 Take Home Message on Translation to Clinic

Translation of a CAD-AI system to the clinic requires an efficient user interface, acceptance testing to validate smooth integration into the workflow and expected performance, adequate user training to ensure proper use and sufficient understanding of CAD-AI performance in the local clinical environment, and robust post-deployment QA procedures to monitor the consistency of performance over time. More advanced validation will involve prospective clinical assessments of the impact of CAD-AI on clinical outcomes using well-designed clinical trial protocols.

7 Discussion

The development of generalizable, robust, and reliable CAD-AI decision support systems is of critical importance for both laboratory proof-of-concept applications and for real-world applications in clinical practice.

To address these important issues, the American Association of Physicists in Medicine (AAPM) assigned a task to the Computer-Aided Image Analysis Subcommittee (CADSC), in part, to develop recommendations on “best practices” for the development, performance assessment, and clinical translation of CAD-AI systems, which are discussed in this task group report. Although we focus on CAD-AI systems for medical imaging, the principles of the processes discussed herein are general and applicable to a broad range of AI applications in the medical field.

A summary of the recommendations (“take home messages”), for best practices for (1) data collection, (2) establishing reference standards, (3) model development, (4) performance assessment, and (5) the translation to clinical practice is presented in Table 3.

Conclusions

The rigor and reproducibility of CAD-AI systems will provide the foundation for the success of such systems when translated into clinical practice. As a community, we are obligated to ensure that the scientific integrity of systems we develop in the laboratory can endure the variabilities and the required reliability in clinical practice to benefit patient care. The topics discussed in this report are all essential elements of CAD-AI systems that, when

diligently considered during system development and validation, should provide the greatest opportunity for successful clinical translation.

Disclosure Statement

The members of AAPM Task Group 273 listed below disclose the following potential Conflict(s) of Interest related to subject matter or materials presented in this document.

Lubomir Hadjiiski - nothing to disclose
Kenny Cha - nothing to disclose
Heang-Ping Chan - nothing to disclose
Karen Drukker - receives royalties from Hologic
Lia Morra - has received funding from HealthTriage srl, not related to this work
Janne J. Näppi - has received royalties from Hologic and from MEDIAN Technologies, through the University of Chicago licensing, not related to this work
Berkman Sahiner - nothing to disclose
Hiroyuki Yoshida - has received royalties from licensing fees to Hologic and Medians Technologies through the University of Chicago licensing, not related to this work
Quan Chen - has received compensations from Carina Medical LLC, not related to this work, provides consulting services for Reflexion Medical, which is unrelated to the content of the TG report
Thomas M. Deserno - nothing to disclose
Hayit Greenspan - nothing to disclose
Henkjan Huisman - has received funding from Siemens Healthineers for a scientific research project, not related to this work
Zhimin Huo - nothing to disclose
Richard Mazurchuk - nothing to disclose
Nicholas Petrick - nothing to disclose
Daniele Regge - nothing to disclose
Ravi Samala - nothing to disclose
Ronald M. Summers - has received royalties from iCAD Medical, ScanMed, Philips, PingAn, Translation Holdings. Lab receives research support from PingAn, not related to this work
Kenji Suzuki - provides consulting services for Canon Medical, which is unrelated to the content of the TG report
Georgia Tourassi - nothing to disclose
Daniel Vergara - nothing to disclose
Samuel G. Armato, III – has received royalties and licensing fees for computer-aided diagnosis through the University of Chicago Consultant, Novartis, not related to this work

Acknowledgments

RMS was supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center.

References

1. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J.R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J.H.F. Rudd, E. Sala, C.-B. Schonlieb, and C. Aix, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, 3, 199-217 (2021).
2. M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P.A. Ioannidis, G.S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," Bmj-British Medical Journal, 368, 1-7 (2020).
3. R. Aggarwal, V. Sounderajah, G. Martin, D.S.W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi, "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," NPJ digital medicine, 4, 65-65 (2021).
4. D.W. Kim, H.Y. Jang, K.W. Kim, Y. Shin, and S.H. Park, "Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers," Korean Journal of Radiology, 20, 405-410 (2019).
5. N. Petrick, B. Sahiner, S.G. Armato, A. Bert, L. Correale, S. Delsanto, M.T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z.M. Huo, Y.L. Jiang, L. Morra, S. Paquerault, V. Raykar, F. Samuelson, R.M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, and H.-P. Chan, "Evaluation of computer-aided detection and diagnosis systems," Medical Physics, 40, 087001 (2013).
6. Z.M. Huo, R.M. Summers, S. Paquerault, J. Lo, J. Hoffmeister, S.G. Armato, M.T. Freedman, J. Lin, S.C.B. Lo, N. Petrick, B. Sahiner, D. Fryd, H. Yoshida, and H.-P. Chan, "Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use," Medical Physics, 40, 077001 (2013).
7. J.F. Cohen, D.A. Korevaar, D.G. Altman, D.E. Bruns, C.A. Gatsonis, L. Hooft, L. Irwig, D. Levine, J.B. Reitsma, H.C.W. de Vet, and P.M.M. Bossuyt, "STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration," Bmj Open, 6, e012799 (2016).
8. J.E. Trost, "Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies," Qualitative Sociology, 9, 54-57 (1986).
9. I. Etikan, S.A. Musa, and R.S. Alkassim, "Comparison of convenience sampling and purposive sampling," American journal of theoretical and applied statistics, 5, 1-4 (2016).
10. I. Pan, S. Agarwal, and D. Merck, "Generalizable Inter-Institutional Classification of Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks," Journal of Digital Imaging, 32, 888-896 (2019).
11. J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, and E.K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," Plos Medicine, 15, e1002683 (2018).
12. X. Feng, M.E. Bernard, T. Hunter, and Q. Chen, "Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation," Physics in Medicine and Biology, 65, 07NT01 (2020).
13. X.X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E.J.

Topol, L.M. Bachmann, P.A. Keane, and A.K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," Lancet Digital Health, 1, E271-E297 (2019).
14. K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, and G.S. Collins, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration," Annals of Internal Medicine, 162, W1-W73 (2015).
15. H.M. Whitney, H. Li, Y. Ji, P. Liu, and M.L. Giger, "Harmonization of radiomic features of breast lesions across international DCE-MRI datasets," Journal of Medical Imaging, 7, 012707 (2020).
16. R.M. Nishikawa, M.L. Giger, K. Doi, C.E. Metz, F.F. Yin, C.J. Vyborny, and R.A. Schmidt, "EFFECT OF CASE SELECTION ON THE PERFORMANCE OF COMPUTER-AIDED DETECTION SCHEMES," Medical Physics, 21, 265-269 (1994).
17. R.M. Nishikawa and L.M. Yarusso, *Variations in measured performance of CAD schemes due to database composition and scoring protocol,* in *Medical Imaging 1998: Image Processing, Pts 1 and 2,* K.M. Hanson, Editor. 1998, p. 840-844.
18. S.G. Armato, R.Y. Roberts, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, G. McLennan, R.M. Engelmann, P.H. Bland, D.R. Aberle, E.A. Kazerooni, H. MacMahon, E.J.R. van Beek, D. Yankelevitz, B.Y. Croft, and L.P. Clarke, "The lung image database consortium (LIDC): Ensuring the integrity of expert-defined "truth"," Academic Radiology, 14, 1455-1463 (2007).
19. K.W. Clark, D.S. Gierada, G. Marquez, S.M. Moore, D.R. Maffitt, J.D. Moulton, M.A. Wolfsberger, P. Koppel, S.R. Phillips, and F.W. Prior, "Collecting 48,000 CT Exams for the Lung Screening Study of the National Lung Screening Trial," Journal of Digital Imaging, 22, 667-680 (2009).
20. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B.D. Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016)," Scientific Data, 6, 6 (2019).
21. "Summary of the HIPAA Privacy Rule," https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html,
22. "International Compilation of Human Research Standards. 2021 Edition. Compiled by Office for Human Research Protections, Office of the Assistant Secretary for Health, U.S. Department of Health and Human Services " https://www.hhs.gov/sites/default/files/ohrp-international-compilation-2021.pdf,
23. H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi, "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation," Ai & Society, 36, 59-77 (2021).
24. M. Gong, S. Wang, L. Wang, C. Liu, J. Wang, Q. Guo, H. Zheng, K. Xie, C. Wang, and Z. Hui, "Evaluation of Privacy Risks of Patients' Data in China: Case Study," Jmir Medical Informatics, 8, (2020).

25. K. Pinhao and M.M. R., *Twenty reasons why GDPR compliance does not exempt companies from adjusting to the LGPD,* in *International Bar Association*2021, https://www.ibanet.org/article/0634B90E-98DE-40E6-953F-2F63CB481F02.
26. D.B. Larson, D.C. Magnus, M.P. Lungren, N.H. Shah, and C.P. Langlotz, "Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework," Radiology, 295, 675-682 (2020).
27. J.R. Geis, A.P. Brady, C.C. Wu, J. Spencer, E. Ranschaert, J.L. Jaremko, S.G. Langer, A.B. Kitts, J. Birch, W.F. Shields, R.V. van Genderen, E. Kotter, J.W. Gichoya, T.S. Cook, M.B. Morgan, A. Tang, N.M. Safdar, and M. Kohli, "Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement," Journal of the American College of Radiology, 16, 1516-1521 (2019).
28. K.Y.E. Aryanto, M. Oudkerk, and P.M.A. van Ooijen, "Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy," European Radiology, 25, 3685-3695 (2015).
29. J.D. Robinson, "Beyond the DICOM Header: Additional Issues in Deidentification," American Journal of Roentgenology, 203, W658-W664 (2014).
30. J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,* in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* A.F. Sorelle and W. Christo, Editors. 2018, PMLR: Proceedings of Machine Learning Research. p. 77-91.
31. Y. Liu, A. Jain, C. Eng, D.H. Way, K. Lee, P. Bui, K. Kanada, G.D. Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G.S. Corrado, L.H. Peng, D.R. Webster, D. Ai, S.J. Huang, Y. Liu, R.C. Dunn, and D. Coz, "A deep learning system for differential diagnosis of skin diseases," Nature Medicine, 26, 900-908 (2020).
32. J.W. Gichoya, I. Banerjee, A.R. Bhimireddy, J.L. Burns, L.A. Celi, L.C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.C. Huang, P.C. Kuo, M.P. Lungren, L.J. Palmer, B.J. Price, S. Purkayastha, A.T. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R.Y. Wang, Z. Zaiman, and H.R. Zhang, "Al recognition of patient race in medical imaging: a modelling study," Lancet Digital Health, 4, E406-E414 (2022).
33. S. Shrestha and S. Das, "Exploring gender biases in ML and AI academic research through systematic literature review," Frontiers in artificial intelligence, 5, 976838-976838 (2022).
34. I. Dankwa-Mullan and D. Weeraratne, "Artificial Intelligence and Machine Learning Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity," Cancer Discovery, 12, 1423-1427 (2022).
35. H.-P. Chan, S.C.B. Lo, B. Sahiner, K.L. Lam, and M.A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Medical Physics, 22, 1555-1567 (1995).
36. A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems,* pp. 1097-1105 (2012).
37. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozairy, A. Courville, and Y. Bengio, "Generative Adversarial Nets," arXiv:1406.2661v1 (2014).
38. M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "SYNTHETIC DATA AUGMENTATION USING GAN FOR IMPROVED LIVER LESION CLASSIFICATION," in *15th IEEE International Symposium on Biomedical Imaging (ISBI),* Washington, DC. pp. 289-293 (2018).

39. K.H. Cha, N. Petrick, A. Pezeshk, C.G. Graff, D. Sharma, A. Badal, and B. Sahiner, "Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning," Journal of medical imaging (Bellingham, Wash.), 7, 012703-012703 (2020).
40. A. Hagiwara, S. Fujita, Y. Ohno, and S. Aoki, "Variability and Standardization of Quantitative Imaging Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence," Investigative Radiology, 55, 601-616 (2020).
41. B. Graham, "Kaggle diabetic retinopathy detection competition report," University of Warwick, (2015).
42. K. Robinson, H. Li, L. Lan, D. Schacht, and M. Giger, "Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM," Medical Physics, 46, 2145-2156 (2019).
43. B. Baessler, K. Weiss, and D.P. dos Santos, "Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging A Phantom Study," Investigative Radiology, 54, 221-228 (2019).
44. S.A. Mali, A. Ibrahim, H.C. Woodruff, V. Andrearczyk, H. Mueller, S. Primakov, Z. Salahuddin, A. Chatterjee, and P. Lambin, "Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods," Journal of Personalized Medicine, 11, (2021).
45. L. Gallardo-Estrella, D.A. Lynch, M. Prokop, D. Stinson, J. Zach, P.F. Judy, B. van Ginneken, and E.M. van Rikxoort, "Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification," European Radiology, 26, 478-486 (2016).
46. M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, and N. Jahanshad, "Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI),* Electr Network. pp. 313-322 (2021).
47. R. Rai, L.C. Holloway, C. Brink, M. Field, R.L. Christiansen, Y. Sun, M.B. Barton, and G.P. Liney, "Multicenter evaluation of MRI-based radiomic features: A phantom study," Medical Physics, 47, 3054-3063 (2020).
48. J.-P. Fortin, D. Parker, B. Tunc, T. Watanabe, M.A. Elliott, K. Ruparel, D.R. Roalf, T.D. Satterthwaite, R.C. Gur, R.E. Gur, R.T. Schultz, R. Verma, and R.T. Shinohara, "Harmonization of multi-site diffusion tensor imaging data," Neuroimage, 161, 149-170 (2017).
49. F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, "Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomics," Radiology, 291, 52-58 (2019).
50. T. Nakahara, H. Daisaki, Y. Yamamoto, T. Iimori, K. Miyagawa, T. Okamoto, Y. Owaki, N. Yada, K. Sawada, R. Tokorodani, and M. Jinzaki, "Use of a digital phantom developed by QIBA for harmonizing SUVs obtained from the state-of-the-art SPECT/CT systems: a multicenter study," Ejnmmi Research, 7, (2017).
51. H. Keller, T. Shek, B. Driscoll, Y. Xu, B. Nghiem, S. Nehmeh, M. Grkovski, C.R. Schmidtlein, M. Budzevich, Y. Balagurunathan, J.J. Sunderland, R.R. Beichel, C. Uribe, T.-Y. Lee, F. Li, D.A. Jaffray, and I. Yeung, "Noise-Based Image Harmonization Significantly Increases Repeatability and Reproducibility of Radiomics Features in PET Images: A Phantom Study," Tomography, 8, 1113-1128 (2022).
52. G. Revesz, H.L. Kundel, and M. Bonitatibus, "THE EFFECT OF VERIFICATION ON THE ASSESSMENT OF IMAGING TECHNIQUES," Investigative Radiology, 18, 194-198 (1983).

53. D.P. Miller, K.F. O'Shaughnessy, S.A. Wood, and R.A. Castellino, *Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions,* in *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment,* D.P. Chakraborty and M.P. Eckstein, Editors. 2004, p. 173-184.
54. Y. Jiang, "A Monte Carlo simulation method to understand expert-panel consensus truth and double readings," Medical Image Perception Conference XII. 2007. The University of Iowa, Iowa City, IA, (2007).
55. S.G. Armato, R.Y. Roberts, M. Kocherginsky, D.R. Aberle, E.A. Kazerooni, H. MacMahon, E.J.R. van Beek, D. Yankelevitz, G. McLennan, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, P. Caligiuri, L.E. Quint, B. Sundaram, B.Y. Croft, and L.P. Clarke, "Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence on the Definition of "Truth"," Academic Radiology, 16, 28-38 (2009).
56. C. Zhou, H.-P. Chan, A. Chughtai, S. Patel, J. Kuriakose, L.M. Hadjiiski, J. Wei, and E.A. Kazerooni, "Variabilities in Reference Standard by Radiologists and Performance Assessment in Detection of Pulmonary Embolism in CT Pulmonary Angiography," Journal of Digital Imaging, 32, 1089-1096 (2019).
57. B. Sahiner, H.-P. Chan, L.M. Hadjiiski, P.N. Cascade, E.A. Kazerooni, A.R. Chughtai, C. Poopat, T. Song, L. Frank, J. Stojanovska, and A. Attili, "Effect of CAD on radiologists' detection of lung nodules on thoracic CT scans: Analysis of an observer performance study by nodule size," Academic Radiology, 1518-1530 (2009).
58. A. Wenzel and H. Hintze, "The choice of gold standard for evaluating tests for caries diagnosis," Dentomaxillofacial Radiology, 28, 132-136 (1999).
59. T.M. Lehmann, *From plastic to gold: A unified classification scheme for reference standards in medical image processing,* in *Medical Imaging 2002: Image Processing, Vol 1-3,* M. Sonka and J.M. Fitzpatrick, Editors. 2002, p. 1819-1827.
60. F. Li, R. Engelmann, S.G. Armato, and H. MacMahon, "Computer-Aided Nodule Detection System: Results in an Unselected Series of Consecutive Chest Radiographs," Academic Radiology, 22, 475-480 (2015).
61. D.F. Yankelevitz and C.I. Henschke, "Does 2-year stability imply that pulmonary nodules are benign?," American Journal of Roentgenology, 168, 325-328 (1997).
62. G.J.S. Litjens, J.O. Barentsz, N. Karssemeijer, and H.J. Huisman, "Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI," European Radiology, 25, 3187-3199 (2015).
63. "DREAM. The digital mammography dream challenge. https://www.synapse.org/Digital_Mammography_DREAM_challenge," (2017).
64. S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F.J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C.J. Kelly, D. King, J.R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J.J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K.C. Young, J. De Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," Nature, 577, 89-94 (2020).
65. C.R. Meyer, T.D. Johnson, G. McLennan, D.R. Aberle, E.A. Kazerooni, H. MacMahon, B.F. Mullan, D.F. Yankelevitz, E.J.R. van Beek, S.G. Armato, M.F. McNitt-Gray, A.P. Reeves, D. Gur, C.I. Henschke, E.A. Hoffman, P.H. Bland, G. Laderach, R. Pais, D. Qing, C. Piker, J.F. Guo, A. Starkey, D. Max, B.Y. Croft, and L.P. Clarke, "Evaluation of lung MDCT nodule annotation across radiologists and methods," Academic Radiology, 13, 1254-1265 (2006).

66. J. Tan, J. Pu, B. Zheng, X. Wang, and J.K. Leader, "Computerized comprehensive data analysis of Lung Imaging Database Consortium (LIDC)," Medical Physics, 37, 3802-3808 (2010).
67. K. Yan, X. Wang, L. Lu, and R.M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," Journal of Medical Imaging, 5, 036501 (2018).
68. L. Oakden-Rayner, "Exploring Large-scale Public Medical Image Datasets," Academic Radiology, 27, 106-112 (2020).
69. D.A. Bluemke, L. Moy, M.A. Bredella, B.B. Ertl-Wagner, K.J. Fowler, V.J. Goh, E.F. Halpern, C.P. Hess, M.L. Schiebler, and C.R. Weiss, "Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the Radiology Editorial Board," Radiology, 294, 487-489 (2020).
70. S. Goel, Y. Sharma, M.-L. Jauer, and T.M. Deserno, "WeLineation: Crowdsourcing delineations for reliable ground truth estimation," Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 11318, 113180C (2020).
71. T.B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J.E. Burns, and R.M. Summers, "Distributed Human Intelligence for Colonic Polyp Classification in Computer-aided Detection for CT Colonography," Radiology, 262, 824-833 (2012).
72. M.-L. Jauer, S. Goel, Y. Sharma, T.M. Deserno, M. Gijs, T.T.J.M. Berendschot, C.J.F. Bertens, and R.M.M.A. Nuijts, "STAPLE performance assessed on crowdsourced sclera segmentations," Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 11318, 113180K (2020).
73. A. Badano, C.G. Graff, A. Badal, D. Sharma, R. Zeng, F.W. Samuelson, S.J. Glick, and K.J. Myers, "Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial," JAMA Network Open, 1, e185474-e185474 (2018).
74. E. Abadi, W.P. Segars, H. Chalian, and E. Samei, "Virtual Imaging Trials for Coronavirus Disease (COVID-19)," American Journal of Roentgenology, 216, 362-368 (2021).
75. R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, and C.D. Richter, "Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis," Physics in Medicine and Biology, 65, 105002 (2020).
76. M. Rajchl, M.C.H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M.A. Rutherford, J.V. Hajnal, B. Kainz, and D. Rueckert, "DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks," Ieee Transactions on Medical Imaging, 36, 674-683 (2017).
77. S.K. Warfield, K.H. Zou, and W.M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," Ieee Transactions on Medical Imaging, 23, 903-921 (2004).
78. N. Petrick, B. Sahiner, H.-P. Chan, M.A. Helvie, S. Paquerault, and L.M. Hadjiiski, "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis - Experience in 263 patients," Radiology, 224, 217-224 (2002).
79. M. Kallergi, G.M. Carney, and J. Gaviria, "Evaluating the performance of detection algorithms in digital mammography," Medical Physics, 26, 267-275 (1999).
80. S.K. Zhou, H. Greenspan, C. Davatzikos, J.S. Duncan, B. Van Ginneken, A. Madabhushi, J.L. Prince, D. Rueckert, and R.M. Summers, "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises," Proceedings of the Ieee, 109, 820-838 (2021).

81. D. Gur, R.F. Wagner, and H.-P. Chan, "On the repeated use of databases for testing incremental improvement of computer-aided detection schemes," Academic Radiology, 11, 103-105 (2004).
82. K. Fukunaga, "Introduction to statistical pattern recognition," 2nd edition. Academic Press, San Diego, (1990).
83. Q. Li and K. Doi, "Comparison of typical evaluation methods for computer-aided diagnostic schemes: Monte Carlo simulation study," Medical Physics, 34, 871-876 (2007).
84. B. Efron, "ESTIMATING THE ERROR RATE OF A PREDICTION RULE - IMPROVEMENT ON CROSS-VALIDATION," Journal of the American Statistical Association, 78, 316-331 (1983).
85. B. Sahiner, H.-P. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited dataset," Medical Physics, 35, 1559-1570 (2008).
86. J.M. Bland and D.G. Altman, "Statistics Notes: Bootstrap resampling methods," Bmj-British Medical Journal, 350, h2622 (2015).
87. R.K. Samala, H.P. Chan, L. Hadjiiski, and M.A. Helvie, "Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification," Medical Physics, 48, 2827-2837 (2021).
88. S. Russell and P. Norving, "Artificial intelligence: a modern approach," 4th Edition, Pearson, USA, (2020).
89. C. Bishop, "Pattern recognition and machine learning," Springer, Singapore, (2006).
90. P. Winston, "Artificial Intelligence," 3rd Edition, Addison-Wesley, USA, (1993).
91. A. Jaiswal, A.R. Babu, M.Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on Contrastive Self-Supervised Learning," Technologies, 9, (2021).
92. J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, "Multi -task contrastive learning for automatic CT and X-ray diagnosis of COVID-19," Pattern Recognition, 114, (2021).
93. J.J. Nappi, R. Tachibana, T. Hironaka, and H. Yoshida, "Electronic cleansing by unpaired contrastive learning in non-cathartic laxative-free CT colonography," Proc SPIE Medical Imaging, 12037, 120370S (2022).
94. N. Tajbakhsh, Y.F. Hu, J.L. Cao, X.J. Yan, Y. Xiao, Y. Lu, J.M. Liang, D. Terzopoulos, and X.W. Ding, *Surrogate Supervision For Medical Image Analysis: Effective Deep Learning From Limited Quantities of Labeled Data,* in *2019 Ieee 16th International Symposium on Biomedical Imaging*. 2019, p. 1251-1255.
95. R. Tachibana, J.J. Nappi, T. Hironaka, and H. Yoshida, "Self-Supervised adversarial learning with a limited dataset for electronic cleansing in computed tomographic colonography: a preliminary feasibility study," Cancers, 14, 4125 (2022).
96. Z. Zhou, V. Sodha, M.M.R. Siddiquee, R. Feng, N. Tajbakhsh, M.B. Gotway, and J. Liang, *Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis,* in *Medical Image Computing and Computer Assisted Intervention - Miccai 2019, Pt Iv,* D. Shen*, et al.*, Editors. 2019, p. 384-393.
97. S. Beiden, G. Campbell, K. Meier, and R. Wagner, "The problem of ROC analysis without truth: the EM algorithm and the information matrix," Proc SPIE Medical Imaging, 3981, 126-134 (2000).
98. V. Cheplygina, M. de Bruijne, and J.P.W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," Medical Image Analysis, 54, 280-296 (2019).
99. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 521, 436-444 (2015).

100. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, *How transferable are features in deep neural networks ?,* in *Advances in Neural Information Processing Systems,* Z. Ghahramani*, et al.*, Editors. 2014, p. 3320-3328.
101. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, *CHEST PATHOLOGY DETECTION USING DEEP LEARNING WITH NON-MEDICAL TRAINING,* in *2015 Ieee 12th International Symposium on Biomedical Imaging*. 2015, p. 294-297.
102. H.C. Shin, H.R. Roth, M.C. Gao, L. Lu, Z.Y. Xu, I. Nogues, J.H. Yao, D. Mollura, and R.M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," Ieee Transactions on Medical Imaging, 35, 1285-1298 (2016).
103. I. Diamant, Y. Bar, O. Geva, L. Wolf, G. Zimmerman, S. Lieberman, E. Konen, and H. Greenspan, *Chest Radiograph Pathology Categorization via Transfer Learning*. Deep Learning for Medical Image Analysis, eds. S.K. Zhou, H. Greenspan, andD. Shen. 2017. 299-320.
104. R.K. Samala, H.-P. Chan, L. Hadjiiski, M.A. Helvie, J. Wei, and K. Cha, "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography," Medical Physics, 43, 6654-6666 (2016).
105. N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, and J.M. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," Ieee Transactions on Medical Imaging, 35, 1299-1312 (2016).
106. J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, and B. Ni, "Reinventing 2D Convolutions for 3D Images," Ieee Journal of Biomedical and Health Informatics, 25, 3009-3018 (2021).
107. R. Tachibana, J.J. Nappi, J. Ota, N. Kohlhase, T. Hironaka, S.H. Kim, D. Regge, and H. Yoshida, "Deep Learning Electronic Cleansing for Single- and Dual-Energy CT Colonography," Radiographics, 38, 2034-2050 (2018).
108. R.K. Samala, H.-P. Chan, L. Hadjiiski, M.A. Helvie, C.D. Richter, and K.H. Cha, "Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning Using Deep Neural Nets," Ieee Transactions on Medical Imaging, 38, 686-696 (2019).
109. X. Mei, Z. Liu, P.M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K.E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z.A. Fayad, and Y. Yang, "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning," Radiology: Artificial Intelligence 4, e210315 (2022).
110. M. Heker and H. Greenspan, "Joint Liver Lesion Segmentation and Classification via Transfer Learning," arXiv preprint arXiv:2004.12352, (2020).
111. R. Caruana, "Multitask learning," In: Learning to learn. Springer, 95-133 (1998).
112. R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, K.H. Cha, and C.D. Richter, "Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms," Physics in Medicine and Biology, 62, 8894-8908 (2017).
113. J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, *Dataset Shift in Machine Learning*. ACM Digital Library, 2009, The MIT Press. 1-248.
114. D.C. Castro, I. Walker, and G. B., "Causality matters in medical imaging," arXiv preprint arXiv:1912.08142, (2019).
115. G. Csurka, *A Comprehensive Survey on Domain Adaptation for Visual Applications,* in *Domain Adaptation in Computer Vision Applications,* G. Csurka, Editor. 2017, Springer: p. 1-35.

116. K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, *Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks,* in *Information Processing in Medical Imaging,* M. Niethammer*, et al.*, Editors. 2017, p. 597-609.
117. A.F. Frangi, S.A. Tsaftaris, and J.L. Prince, "Simulation and Synthesis in Medical Imaging," Ieee Transactions on Medical Imaging, 37, 673-679 (2018).
118. F. Mahmood, R. Chen, and N.J. Durr, "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training," Ieee Transactions on Medical Imaging, 37, 2572-2581 (2018).
119. H.C. Shin, N.A. Tenenholtz, J.K. Rogers, C.G. Schwarz, M.L. Senjem, J.L. Gunter, K.P. Andriole, and M. Michalski, *Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks,* in *Simulation and Synthesis in Medical Imaging,* A. Gooya*, et al.*, Editors. 2018, p. 1-11.
120. V. Sandfort, K. Yan, P.J. Pickhardt, and R.M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," Scientific Reports, 9, 16884 (2019).
121. H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A.Y. Arcas, *Communication-Efficient Learning of Deep Networks from Decentralized Data,* in *Artificial Intelligence and Statistics, Vol 54,* A. Singh and J. Zhu, Editors. 2017, p. 1273-1282.
122. J. Konecny, H.B. McMahan, F.X. Yu, P. Richtarik, A.T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, (2016).
123. K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," Journal of the American Medical Informatics Association, 25, 945-954 (2018).
124. N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R.M. Summers, A. Trask, D. Xu, M. Baust, and M.J. Cardoso, "The future of digital health with federated learning," Npj Digital Medicine, 3, 119 (2020).
125. X. Li, Y. Gu, N. Dvornek, L.H. Staib, P. Ventola, and J.S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," Medical Image Analysis, 65, 101765 (2020).
126. P. McClure, C.Y. Zheng, J.R. Kaczmarzyk, J.A. Lee, S.S. Ghosh, D. Nielson, P. Bandettini, and F. Pereira, *Distributed Weight Consolidation: A Brain Segmentation Case Study,* in *Advances in Neural Information Processing Systems 31,* S. Bengio*, et al.*, Editors. 2018, p. 4093-4103.
127. P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, , and R.G. d'Oliveira, "Advances and open problems in federated learning," arXiv preprint arXiv:1912.04977, (2019).
128. S. Grossberg, "Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world," Neural Networks, 37, 1-47 (2013).
129. G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," Neural Networks, 113, 54-71 (2019).
130. R.M. French, "Catastrophic forgetting in connectionist networks," Trends in Cognitive Sciences, 3, 128-135 (1999).
131. I.J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," arXiv preprint arXiv:1312.6211, (2013).

132. C.E. Metz, "ROC METHODOLOGY IN RADIOLOGIC IMAGING," Investigative Radiology, 21, 720-733 (1986).
133. D.P. Chakraborty and L.H.L. Winter, "FREE-RESPONSE METHODOLOGY - ALTERNATE ANALYSIS AND A NEW OBSERVER-PERFORMANCE EXPERIMENT," Radiology, 174, 873-881 (1990).
134. B.D. Gallas, H.-P. Chan, C.J. D'Orsi, L.E. Dodd, M.L. Giger, D. Gur, E.A. Krupinski, C.E. Metz, K.J. Myers, N.A. Obuchowski, B. Sahiner, A.Y. Toledano, and M.L. Zuley, "Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA," Academic Radiology, 19, 463-477 (2012).
135. K. Doi, H. MacMahon, S. Katsuragawa, R.M. Nishikawa, and Y.L. Jiang, "Computer-aided diagnosis in radiology: potential and pitfalls," European Journal of Radiology, 31, 97-109 (1999).
136. F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, and R.A. Rosati, "EVALUATING THE YIELD OF MEDICAL TESTS," Jama-Journal of the American Medical Association, 247, 2543-2546 (1982).
137. F.E. Harrell, K.L. Lee, and D.B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," Statistics in Medicine, 15, 361-387 (1996).
138. N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," Cancer Chemotherap Rep, 50, 163-170 (1966).
139. P. Therasse, S.G. Arbuck, E.A. Eisenhauer, J. Wanders, R.S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A.T. van Oosterom, M.C. Christian, and S.G. Gwyther, "New Guidelines to Evaluate the Response to Treatment in Solid Tumors," JNCI: Journal of the National Cancer Institute, 92, 205-216 (2000).
140. E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," European Journal of Cancer, 45, 228-247 (2009).
141. P.R. Cohen, "Empirical methods for artificial intelligence," The MIT Press, (1995).
142. X.-H. Zhou, N.A. Obuchowski, and D.K. McClish, "Statistical methods in diagnostic medicine. Wiley; New York," (2002).
143. P. Schober, S.M. Bossers, and L.A. Schwarte, "Statistical Significance Versus Clinical Importance of Observed Effect Sizes: What Do P Values and Confidence Intervals Really Represent?," Anesthesia and Analgesia, 126, 1068-1072 (2018).
144. S.N. Goodman, "Toward evidence-based medical statistics. 1: The P value fallacy," Annals of Internal Medicine, 130, 995-1004 (1999).
145. M. Aickin and H. Gensler, "Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods," American Journal of Public Health, 86, 726-728 (1996).
146. P. Rajpurkar, J. Irvin, R.L. Ball, K.L. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C.P. Langlotz, B.N. Patel, K.W. Yeom, K. Shpanskaya, F.G. Blankenberg, J. Seekins, T.J. Amrhein, D.A. Mong, S.S. Halabi, E.J. Zucker, A.Y. Ng, and M.P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," Plos Medicine, 15, e1002686 (2018).
147. A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 542, 115-118 (2017).

148. A. Rodriguez-Ruiz, K. Lang, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T.H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M.G. Wallis, I. Andersson, S. Zackrisson, R.M. Mann, and I. Sechopoulos, "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists," Jnci-Journal of the National Cancer Institute, 111, 916-922 (2019).
149. H.P. Chan, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, K.L. Lam, T. Ogura, Y.Z. Wu, and H. Macmahon, "IMPROVEMENT IN RADIOLOGISTS DETECTION OF CLUSTERED MICROCALCIFICATIONS ON MAMMOGRAMS - THE POTENTIAL OF COMPUTER-AIDED DIAGNOSIS," Investigative Radiology, 25, 1102-1110 (1990).
150. L.M. Hadjiiski, H.-P. Chan, B. Sahiner, M.A. Helvie, M. Roubidoux, C. Blane, C. Paramagul, N. Petrick, J. Bailey, K. Klein, M. Foster, S. Patterson, D. Adler, A. Nees, and J. Shen, "Breast Masses: Computer-aided Diagnosis with Serial Mammograms," Radiology, 240, 343-356 (2006).
151. S.V. Beiden, R.F. Wagner, K. Doi, R.M. Nishikawa, M. Freedman, S.C. Ben Lo, and X.W. Xu, "Independent versus sequential reading in ROC studies of computer-assist modalities: Analysis of components of variance," Academic Radiology, 9, 1036-1043 (2002).
152. C.E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Investigative Radiology, 24, 234-245 (1989).
153. "U.S. Food and Drug Adminstration. Guidance for industry and FDA staff: Computer-assisted detection devices applied to radiology images and radiology device data – premarket notification [510(k)] submissions. 2012 Nov. 21, 2017]; ," Available from: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187294.pdf, (2017).
154. "U.S. Food and Drug Adminstration. Guidance for industry and FDA staff: 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. 2012 Nov. 21, 2017]; Available from: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187315.pdf.," (2017).
155. F.W. Samuelson and C.K. Abbey, "The Reproducibility of Changes in Diagnostic Figures of Merit Across Laboratory and Clinical Imaging Reader Studies," Academic Radiology, 24, 1436-1446 (2017).
156. B.D. Gallas, W. Chen, E. Cole, R. Ochs, N. Petrick, E.D. Pisano, B. Sahiner, F.W. Samuelson, and K.J. Myers, "Impact of prevalence and case distribution in lab-based diagnostic imaging studies," Journal of Medical Imaging, 6, 015501 (2019).
157. R.F. Wagner, C.E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: A tutorial review," Academic Radiology, 14, 723-748 (2007).
158. N.A. Obuchowski, "New methodological tools for multiple-reader ROC studies," Radiology, 243, 10-12 (2007).
159. H.-P. Chan, B. Sahiner, R.F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Medical Physics, 26, 2654-2668 (1999).
160. B. Sahiner, H.-P. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," Medical Physics, 27, 1509-1522 (2000).

161. X. Bouthillier, C. Laurent, and P. Vincent, "Unreproducible research is reproducible," in *36th International Conference on Machine Learning, ICML 2019,* pp. 1150-1159 (2019).
162. M. McDermott, S. Wang, Marinsek, N. Ranganath, R. Ghassemi, and L. M. Foschini, "Reproducibility in machine learning for health," arXiv preprint arXiv:1907.01463, (2019).
163. S.N. Goodman, D. Fanelli, and J.P.A. Ioannidis, "What does research reproducibility mean?," Science Translational Medicine, 8, 341ps12 (2016).
164. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *33rd International Conference on Machine Learning, ICML 2016,* pp. 1651-1660 (2016).
165. A. Kendall and Y. Gal, *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?,* in *Advances in Neural Information Processing Systems 30,* I. Guyon*, et al.*, Editors. 2017,
166. R. Robinson, V.V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M.M. Sanghvi, N. Aung, J.M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A.M. Lee, V. Carapella, Y.J. Kim, S.K. Piechnik, S. Neubauer, S.E. Petersen, C. Page, P.M. Matthews, D. Rueckert, and B. Glocker, "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study," Journal of Cardiovascular Magnetic Resonance, 21, (2019).
167. Y. Yang, X. Guo, Y. Pan, P. Shi, H. Lv, and T. Ma, "Uncertainty Quantification in Medical Image Segmentation with Multi-decoder U-Net," arXiv preprint arXiv:2109.07045, (2021).
168. M. Rezaei, J. Näppi, B. Bischl, and H. Yoshida, "Bayesian uncertainty estimation for detection of long-tail and unseen conditions in abdominal images," Proc of SPIE Medical Imaging, 12033, 1203311 (2022).
169. Z. Salahuddin, H.C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," Computers in Biology and Medicine, 140, (2022).
170. M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R.M. Summers, and R. Wiest, "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities," Radiology. Artificial intelligence, 2, e190043-e190043 (2020).
171. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Mueller, "Evaluating the Visualization of What a Deep Neural Network Has Learned," Ieee Transactions on Neural Networks and Learning Systems, 28, 2660-2673 (2017).
172. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning Deep Features for Discriminative Localization,* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. 2016, p. 2921-2929.
173. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,* in *2017 Ieee International Conference on Computer Vision*. 2017, p. 618-626.
174. H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* pp. 111-119 (2020).
175. A.J. Barnett, F.R. Schwartz, C. Tao, C. Chen, Y. Ren, J.Y. Lo, and C. Rudin, "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," Nature Machine Intelligence, 3, 1061-+ (2021).

176. N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M.D. Li, and J. Kalpathy-Cramer, "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging," Radiology. Artificial intelligence, 3, e200267-e200267 (2021).
177. M.T. Ribeiro, S. Singh, C. Guestrin, and M. Assoc Comp, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016. 1135-1144.
178. H.P. Chan, R.K. Samala, L.M. Hadjiiski, and C. Zhou, *Deep Learning in Medical Image Analysis,* in *Deep Learning in Medical Image Analysis: Challenges and Applications,* G. Lee and H. Fujita, Editors. 2020, p. 3-21.
179. H.-P. Chan, L.M. Hadjiiski, and R.K. Samala, "Computer-aided diagnosis in the era of deep learning," Medical Physics, 47, e218-e227 (2020).
180. T.W. Freer and M.J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology, 220, 781-786 (2001).
181. M.A. Helvie, L. Hadjiiski, E. Makariou, H.-P. Chan, N. Petrick, B. Sahiner, S.C.B. Lo, M. Freedman, D. Adler, J. Bailey, C. Blane, D. Hoff, K. Hunt, L. Joynt, K. Klein, C. Paramagul, S.K. Patterson, and M.A. Roubidoux, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial," Radiology, 231, 208-214 (2004).
182. R.L. Birdwell, P. Bandodkar, and D.M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," Radiology, 236, 451-457 (2005).
183. J.C. Dean and C.C. Ilvento, "Improved cancer detection using computer-aided detection with diagnostic and screening mammography: Prospective study of 104 cancers," American Journal of Roentgenology, 187, 20-28 (2006).
184. M.J. Morton, D.H. Whaley, K.R. Brandt, and K.K. Amrami, "Screening mammograms: Interpretation with computer-aided detection - Prospective evaluation," Radiology, 239, 375-383 (2006).
185. F.J. Gilbert, S.M. Astley, M.G. Gillan, O.F. Agbaje, M.G. Wallis, J. James, C.R. Boggis, S.W. Duffy, and C.I. Grp, "CADET II: A prospective trial of computer-aided detection (CAD) in the UK Breast Screening Programme," Journal of Clinical Oncology, 26, 508 (2008).
186. D. Regge, P. Della Monica, G. Galatola, C. Laudi, A. Zambon, L. Correale, R. Asnaghi, B. Barbaro, C. Borghi, D. Campanella, M.C. Cassinis, R. Ferrari, A. Ferraris, R. Golfieri, C. Hassan, F. Iafrate, G. Iussich, A. Laghi, R. Massara, E. Neri, L. Sali, S. Venturini, and G. Gandini, "Efficacy of Computer-aided Detection as a Second Reader for 6-9-mm Lesions at CT Colonography: Multicenter Prospective Trial," Radiology, 266, 168-176 (2013).
187. J. Concato, N. Shah, and R.I. Horwitz, "Randomized, controlled trials, observational studies, and the hierarchy of research designs," New England Journal of Medicine, 342, 1887-1892 (2000).
188. D. Gur, J.H. Sumkin, H.E. Rockette, M. Ganott, C. Hakim, L. Hardesty, W.R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," Journal of the National Cancer Institute, 96, 185-190 (2004).
189. J.J. Fenton, L. Abraham, S.H. Taplin, B.M. Geller, P.A. Carney, C. D'Orsi, J.G. Elmore, W.E. Barlow, and C. Breast Canc Surveillance, "Effectiveness of Computer-

Aided Detection in Community Mammography Practice," Journal of the National Cancer Institute, 103, 1152-1161 (2011).
190. M. Gromet, "Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms," American Journal of Roentgenology, 190, 854-859 (2008).
191. C.D. Lehman, R.D. Wellman, D.S.M. Buist, K. Kerlikowske, A.N.A. Tosteson, D.L. Miglioretti, and S. Breast Canc, "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection," Jama Internal Medicine, 175, 1828-1837 (2015).
192. S. Cruz Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, A.I. Spirit, C.-A.W. Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, and C.-A.C. Group, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Nature medicine, 26, 1351-1363 (2020).
193. X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.I. Spirit, and C.-A.W. Group, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nature medicine, 26, 1364-1374 (2020).

Figure Legends

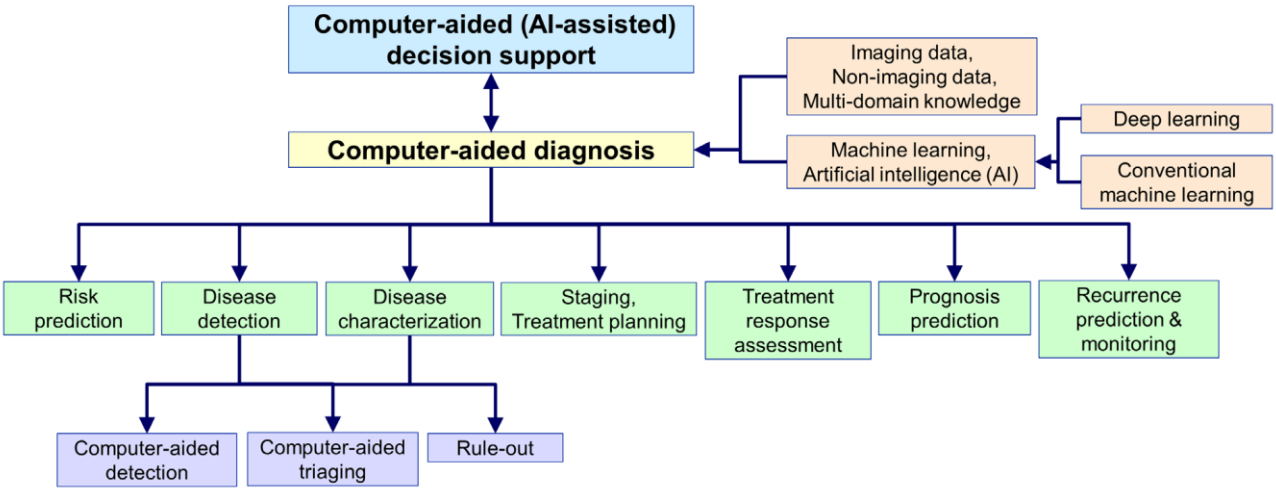

Figure 1. Overview of computer-aided diagnosis applications

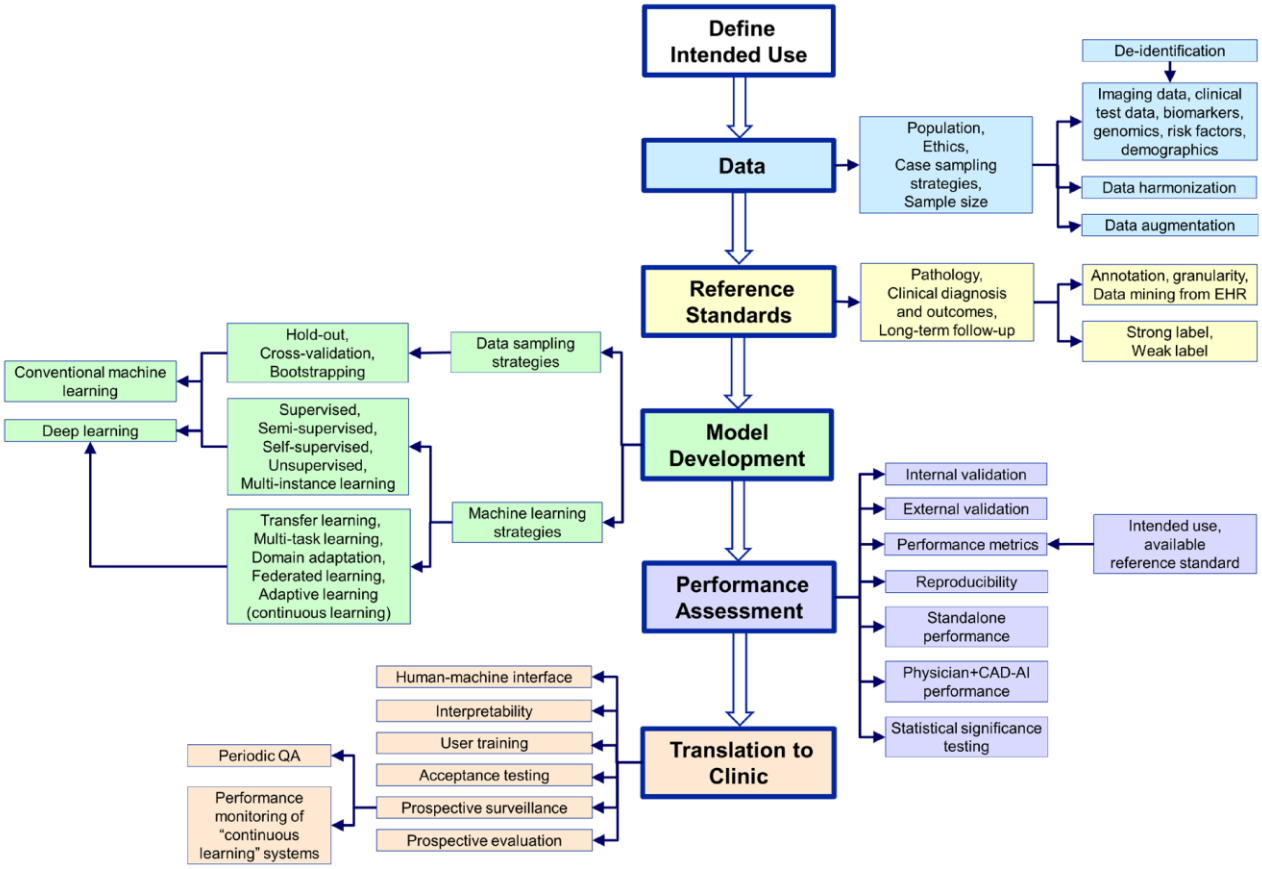

Figure 2. Overview of development of computer-aided decision support systems

Table 1. Type of data shift.

Data Shift	**Definition**
Prevalence shift	training and test datasets have different disease prevalence (class imbalance)
Acquisition or domain shift	different imaging equipment or imaging protocols are used between training and test datasets
Population shift	intrinsic characteristics (e.g., demographics or disease presentation) of the populations under study differ between training and test datasets
Annotation or label shift	class definition changes between training and test datasets, e.g., due to inter-rater variability or lack of standardization in the class definitions

Table 2. Different paradigms of CAD-AI systems.

Paradigm	Intended Use	Evaluation approach
Second read	Improving decision making by providing a second opinion to the physician *after* initial interpretation	Assessment of physician performance without and with the aid in a sequential reader study design; first interpret each case without, then with, CAD-AI system [5, 134, 149-151]; or independent or crossover study design similar to that of concurrent read.
Concurrent read	Improving decision making by showing system output to the physician *at the same time* as initial interpretation	Assessment of physician performance without and with the aid in an independent or crossover reader study design; cases are interpreted in batches either with or without the aid after a sufficient washout time and in counter-balanced manner to reduce the potential memorization effects [5, 134, 152]
Triage	Improving workflow by prioritization: All cases are interpreted but order prioritized by CAD-AI system [153, 154]	Assessment of process improvement by local clinical operations
Rule-out	Improving workflow by removal of normal or negative cases from workflow: The removed cases are not interpreted by physician.	Comparison of performance with and without rule-out in clinical practice [146-148]

Table 3. Summary of recommendations on the best practices and standards for the development and performance assessment of computer-aided decision support systems.

Section	Take Home Message
Data	In summary, proper data collection methods are of critical importance to successful training, validation, and implementation of CAD-AI algorithms. Improper collection and manipulation of data (such as improper data augmentation) can lead to an overestimation of performance or lack of generalizability.
Reference Standards	The required type and granularity of the reference standard depends on the task at hand. An objective reference standard is preferred; however, when a subjective reference standard cannot be avoided, independent assessments of multiple domain experts

	should be obtained and their variabilities should be evaluated.
Model Development	Training approaches, especially for deep learning algorithms, are continuously improving with the goal of achieving robust, effective, and privacy-preserving CAD-AI models. An independent test set representative of the intended use that was not employed to guide model optimization in any learning paradigm is of critical importance. Robust training methods, although important for all CAD-AI systems, are especially important for systems that may operate in clinical practice with minimal or no human supervision.
Performance Assessment	The most appropriate performance metric(s) will depend on the task and the reference standard. Often multiple performance metrics are appropriate and use of multiple metrics is frequently desirable. Power calculations should be an integral part of study design, and performance analysis should include error estimates, assessment of statistical significance, and preferably assessment of reproducibility.
Translation to Clinic	Translation of a CAD-AI system to the clinic requires an efficient user interface, acceptance testing to validate smooth integration into the workflow and expected performance, adequate user training to ensure proper use and sufficient understanding of CAD-AI performance in the local clinical environment, and robust post-deployment QA procedures to monitor the consistency of performance over time. More advanced validation will involve prospective clinical assessments of the impact of CAD-AI on clinical outcomes using well-designed clinical trial protocols.