Nuclear Medicine and Artificial Intelligence: Best Practices for Evaluation (the RELAINCE Guidelines)

Abhinav K. Jha[1], Tyler J. Bradshaw[2], Irène Buvat[3], Mathieu Hatt[4], Prabhat KC[5], Chi Liu[6], Nancy F. Obuchowski[7], Babak Saboury[8], Piotr J. Slomka[9], John J. Sunderland[10], Richard L. Wahl[11], Zitong Yu[12], Sven Zuehlsdorff[13], Arman Rahmim[14], and Ronald Boellaard[15]

[1]*Department of Biomedical Engineering and Mallinckrodt Institute of Radiology, Washington University in St. Louis, Missouri;* [2]*Department of Radiology, University of Wisconsin-Madison, Madison, Wisconsin;* [3]*LITO, Institut Curie, Université PSL, U1288 Inserm, Orsay, France;* [4]*LaTiM, INSERM, UMR 1101, Univ Brest, Brest, France;* [5]*Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland;* [6]*Department of Radiology and Biomedical Imaging, Yale University, Connecticut;* [7]*Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio;* [8]*Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Maryland;* [9]*Department of Imaging, Medicine, and Cardiology, Cedars-Sinai Medical Center, California;* [10]*Departments of Radiology and Physics, University of Iowa, Iowa;* [11]*Mallinckrodt Institute of Radiology, Washington University in St. Louis, Missouri;* [12]*Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri;* [13]*Siemens Medical Solutions USA, Inc., Hoffman Estates, Illinois;* [14]*Departments of Radiology and Physics, University of British Columbia, Canada; and* [15]*Department of Radiology & Nuclear Medicine, Cancer Centre Amsterdam, Amsterdam University Medical Centers, Netherlands*

An important need exists for strategies to perform rigorous objective clinical-task-based evaluation of artificial intelligence (AI) algorithms for nuclear medicine. To address this need, we propose a 4-class framework to evaluate AI algorithms for promise, technical task-specific efficacy, clinical decision making, and postdeployment efficacy. We provide best practices to evaluate AI algorithms for each of these classes. Each class of evaluation yields a claim that provides a descriptive performance of the AI algorithm. Key best practices are tabulated as the RELAINCE (Recommendations for EvaLuation of AI for NuClear medicinE) guidelines. The report was prepared by the Society of Nuclear Medicine and Molecular Imaging AI Task Force Evaluation team, which consisted of nuclear-medicine physicians, physicists, computational imaging scientists, and representatives from industry and regulatory agencies.

Key Words: artificial intelligence; evaluation; best practices; clinical task; PET; SPECT; technical efficacy; clinical decision making; postdeployment; generalizability

J Nucl Med 2022; 63:1288–1299
DOI: 10.2967/jnumed.121.263239

I. INTRODUCTION

Artificial intelligence (AI)–based algorithms are showing tremendous promise across multiple aspects of nuclear medicine, including image acquisition, reconstruction, postprocessing, segmentation, diagnostics, and prognostics. Translating this promise to clinical reality requires rigorous evaluations of these algorithms. Insufficient evaluation of AI algorithms may have multiple adverse consequences, including reducing credibility of research findings, misdirection of future research, and, most importantly, yielding tools that are useless or even harmful to patients (*1*). The goal of this report is to provide best practices to evaluate AI algorithms developed for different parts of the imaging pipeline ranging from image acquisition to postprocessing to clinical decision making in the context of nuclear medicine. We provide these practices in the context of evaluating AI algorithms that use artificial neural network–based architectures, including deep learning. However, many principles are broadly applicable to other machine-learning and physics-based algorithms. In the rest of the report, AI algorithms refer to those that use artificial neural networks.

Evaluation has a well-established and essential role in the translation of any imaging technology but is even more critical for AI algorithms due to their working principles. AI algorithms are typically not programmed with user-defined rules, but instead learn rules via analysis of training data. These rules are often not explicit and thus not easily interpretable, leading to unpredictability in output. This leads to multiple unique challenges. First, AI algorithms may yield inaccurate results that may adversely impact performance on clinical tasks. For example, AI-based reconstruction may introduce spurious lesions (*2*), AI-based denoising may remove lesions (*3*), and AI-based lesion segmentation may incorrectly identify healthy tissue as malignancies (*4*). Evaluations are thus crucial to assess the algorithm's clinical utility. A second challenge is that of generalizability. AI algorithms are often complicated models with many tunable parameters. These algorithms may perform well on training data, but may not generalize to new data, such as from a different institution (*5*), population groups (*6,7*), or scanners (*8*). Possible reasons for this include that the algorithm uses data features that correlate with the target outcome only within training data, or that the training data does not sufficiently represent the patient population. Evaluations are needed to assess the generalizability of these algorithms. A third challenge is data drift during clinical deployment. When using AI systems clinically, over time, the input-data distribution may drift from that of the training data due to changes in patient demographics, hardware, acquisition and analysis protocols (*9*). Evaluation in postdeployment settings can help identify this data drift. Rigorous evaluation of AI algorithms is also necessary because AI is being explored to support decisions in high-risk applications, such as guiding treatment.

Received Sep. 17, 2021; revision accepted May 11, 2022.
For correspondence or reprints, contact Abhinav K. Jha (a.jha@wustl.edu).
Published online May 26, 2022
COPYRIGHT © 2022 by the Society of Nuclear Medicine and Molecular Imaging.

In summary, there is an important need for carefully defined strategies to evaluate AI algorithms, and such strategies should be able to address the unique challenges associated with AI techniques. To address this need, the Society of Nuclear Medicine and Molecular Imaging put together an Evaluation team within the AI Task Force. The team consisted of computational imaging scientists, nuclear medicine physicians, nuclear medicine physicists, biostatisticians, and representatives from industry and regulatory agencies. The team was tasked with defining best practices for evaluating AI algorithms for nuclear medicine imaging. This report has been prepared by this team.

In medical imaging, images are acquired for specific clinical tasks. Thus, AI algorithms developed for the various parts of the imaging pipeline, including acquisition, reconstruction, postprocessing, and segmentation, should be evaluated on the basis on how well they assist in the clinical tasks. As described later, these tasks can be broadly classified into 3 categories: classification, quantification, or a combination of both (*10,11*). An oncologic PET image may be acquired for the task of tumor-stage classification or for quantification of tracer uptake in tumor. However, current Al-algorithm evaluation strategies are often task agnostic. For example, AI algorithms for reconstruction and postprocessing are often evaluated by measuring image fidelity to a reference standard using figures of merit (FoMs) such as root mean square error. Similarly, AI-based segmentation algorithms are evaluated using FoMs such as Dice scores. However, studies, including recent ones, show that these evaluation strategies may not correlate with clinical-task performance and task-based evaluations may be needed (*2,3,11–15*). One study observed that evaluation of a reconstruction algorithm for whole-body FDG PET using fidelity-based FoMs indicated excellent performance, but on the lesion-detection task, the algorithm was yielding both false-negatives and -positives due to blurring and pseudo-low uptake patterns, respectively (*2*). Similarly, an AI-based denoising method for cardiac SPECT studied using realistic simulations seemed to yield excellent performance as evaluated using fidelity-based FoMs. However, on the task of detecting perfusion defects, no performance improvement was observed compared with noisy images (*3*). Such findings show that task-agnostic approaches to evaluate AI algorithms have crucial limitations in quantifying performance on clinical tasks. Thus, evaluation strategies that specifically measure performance on clinical tasks are needed.

Evaluation studies should also quantitatively describe the generalizability of the AI algorithm to different population groups and to different portions of the imaging pipeline, including scanners, acquisition, and analysis protocols. Finally, evaluations should yield quantitative measures of performance to enable clear comparison with standard of care and other methods and provide guidance for clinical utility. To incorporate these needs, we recommend that an AI-algorithm evaluation strategy should always produce a claim consisting of the following components (Fig. 1):

- A clear definition of the task
- Patient population(s) for whom the task is defined
- Definition of the imaging process (acquisition, reconstruction, and analysis protocols)
- Process to extract task-specific information
- FoM to quantify task performance, including process to define reference standard

We describe each component in the next section. We next propose an evaluation framework that categorizes the evaluation strategies into 4 classes: proof of concept, technical, clinical and postdeployment evaluation. This framework will serve as a guide to conduct the evaluation study that provides evidence to support the intended claim. We also provide best practices for conducting evaluations for each class. Key best practices are summarized as the RELAINCE (Recommendations for EvaLuation of AI for NuClear medicinE) guidelines.

In this report, the terms "training," "validation," and "testing" will denote the building of a model on a specific dataset, the tuning/optimization of the model parameters, and the evaluation of the optimized model, respectively. The focus of this report is purely on testing/evaluation of an already developed AI algorithm. Best practices for development of AI algorithms are described in a companion paper (*16*).

NOTEWORTHY

- AI algorithms should be evaluated on clinical tasks.
- AI algorithm evaluations should yield a claim that provides a clear and descriptive characterization of the performance of the AI algorithm on a clinical task. The claim should include a definition of the clinical task, patient population for whom the task is defined, definition of the imaging process, procedure to extract task-specific information, and figure of merit to quantify task performance.
- We propose a 4-class framework that evaluates AI algorithms for nuclear-medicine imaging on clinical tasks and yields a claim. The 4 classes in the framework include promise, technical, clinical, and postdeployment evaluation of AI algorithms.
- We provide best practices for determining study type, data collection, defining reference standard, and choosing figures of merit for each class of evaluation.
- Key recommendations are summarized as the RELAINCE (Recommendations for EvaLuation of AI for NuClear medicinE) guidelines.

II. COMPONENTS OF THE CLAIM

The claim provides a clear and descriptive characterization of the performance of an AI algorithm based on how well it assists in the clinical task. The components of a claim are shown in Figure 1 and described below.

II.1. Definition of the Clinical Task

In this paper, the term "task" refers to the clinical goal for which the image was acquired. Broadly, in nuclear medicine, tasks can be grouped into 3 categories: classification (including lesion detection), quantification, or joint classification and quantification. A classification task is defined as one where the patient image is used to classify the patient into one of several categories. For example, identifying if cancer is present or absent or the cancer stage from an oncologic PET image. Similarly, predicting whether a patient would/would not respond to therapy would be a classification task. A quantification task is defined as one where some numeric or statistical feature is estimated from the patient image. Examples include quantifying SUV, metabolic tumor volume (MTV), intralesion heterogeneity, or kinetic parameters from oncologic PET images.

II.2. Patient Population for Whom the Task Is Defined

The performance of an imaging algorithm can be affected by the physical and statistical properties of the imaged patient

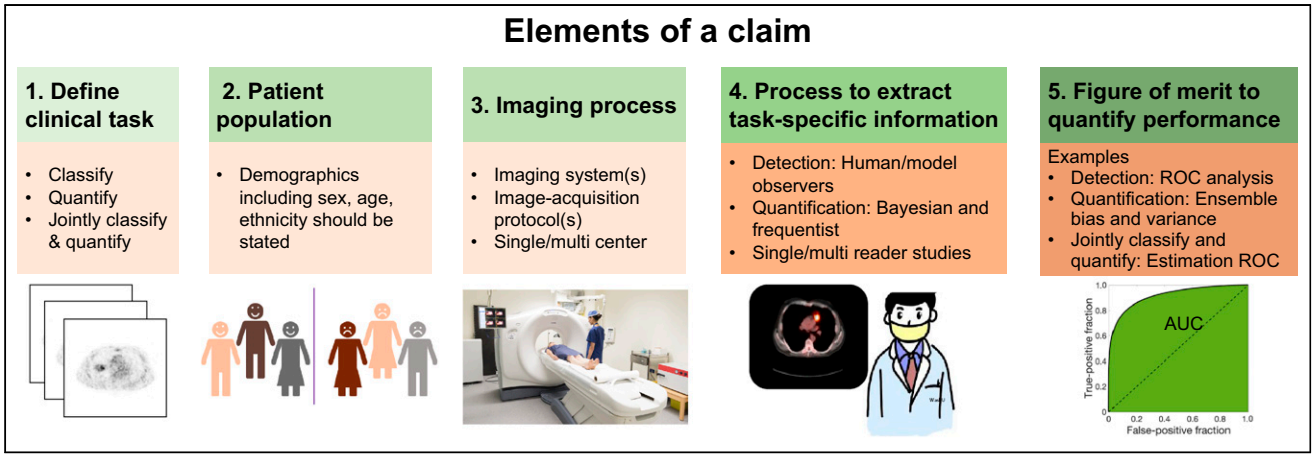

FIGURE 1. The components of a claim. (Scanner image: iStock photo.)

population. Results for one population may not necessarily translate to others (*5,7*). Thus, the patient population should be defined in the claim. This includes aspects such as sex, ethnicity, age group, geographic location, disease stage, social determinants of health, and other disease and application-relevant biologic variables.

II.3. Definition of Imaging Process

The imaging system, acquisition protocol, and reconstruction and analysis parameters may affect task performance. For example, an AI algorithm evaluated for a high-resolution PET system may rely on high-frequency features captured by this system and thus not apply to low-resolution systems (*8*). Depending on the algorithm, specific acquisition protocol parameters may need to be specified or the requirement to comply with a certain accreditation standard, such as SNMMI-Clinical Trial Network, RSNA QIBA profile, and the EARL standards, may need to be stated. For example, an AI-based denoising algorithm for ordered-subsets-expectation-maximization (OSEM)–based reconstructed images may not apply to images reconstructed using filtered backprojection or even for a different number of OSEM iterations since noise properties change with iteration numbers. Thus, depending on the application, the claim should specify the imaging protocol. Further, if the algorithm was evaluated across multiple scanners, or with multiple protocols, that should be specified.

II.4. Process to Extract Task-Specific Information

Task-based evaluation of an imaging algorithm requires a strategy to extract task-specific information from the images. For classification tasks, a typical strategy is to have human observer(s) read the images, detect lesions, and classify the patient or each detected lesion into a certain class (e.g., malignant or benign). Here, observer competency (multiple trained radiologists/one trained radiologist/resident/untrained reader) will impact task performance. The choice of the strategy may impact confidence of the validity of the algorithm. This is also true for quantification and joint classification/quantification tasks. Thus, this strategy should be specified in the claim.

II.5. Figure of Merit (FoM) to Quantify Task Performance

FoMs quantitatively describe the algorithm's performance on the clinical task, enabling comparison of different methods, comparison to standard of care, and defining quantitative metrics of success. FoMs should be accompanied by confidence intervals (CIs), which quantify uncertainty in performance. To obtain the FoM, a reference standard is needed. The process to define the reference standard should be stated.

The Claim Describes the Generalizability of an AI Algorithm: Generalizability is defined as an algorithm's ability to properly work with new, previously unseen data, such as that from a different institution, scanner, acquired with a different image-acquisition protocol, or processed by a different reader. By providing all the components of a claim, an evaluation study will describe the algorithm's generalizability to unseen data, since the claim will specify the characteristics of the population used for evaluation, state whether the evaluation was single or multicenter, define the image acquisition and analysis protocols used, as well as the competency of the observer performing the evaluation study. Figure 2 presents a schematic showing how different kinds of generalizability could be established. Some key points from this figure are:

- Providing evidence for generalizability requires external validation. This is defined as validation where some portion of the testing study, such as the data (patient population demographics) or the process to acquire the data, is different from that in the development cohort. Depending on the level of external validation, the claim can be appropriately defined.
- For a study that claims to be generalizable across populations, scanners, and readers, the external cohort would be from different patient demographics, with different scanners, and analyzed by different readers than the development cohort, respectively.
- Multicenter studies provide higher confidence about generalizability compared with single-center studies since they typically include some level of external validation (patients from different geographical locations/different scanners/different readers).

III. METHODS FOR EVALUATION

The evaluation framework for AI algorithms is provided in Figure 3. The 4 classes of this framework are differentiated based on their objectives, as briefly described below, with details provided in the ensuing subsections. An example for an AI low-dose PET reconstruction algorithm is provided. Figure 3 contains another example for an AI-based automated segmentation algorithm. A detailed example of using this framework to evaluate a hypothetical AI-based transmission-less attenuation compensation method for SPECT

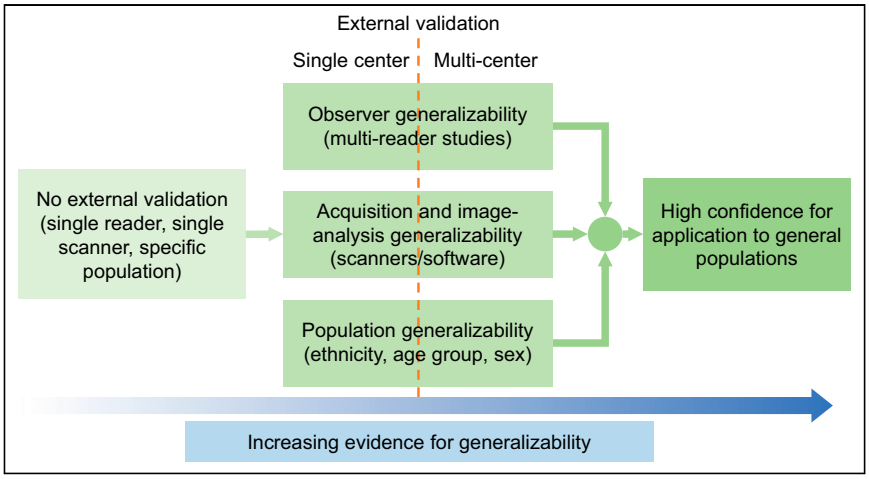

FIGURE 2. Increasing levels of rigor of evaluation, and how they in turn provide increased confidence in the generalizability.

(Supplemental Fig. 1; supplemental materials are available at http://jnm.snmjournals.org) (*17*) is provided in Supplemental section A.

- Class 1: Proof-of-concept (POC) evaluation: Shows the novelty and promise of an algorithm proposed using task-agnostic FoMs. Provides promise for further clinical task-specific evaluation.
  Example: Evaluating the AI PET reconstruction algorithm using root mean square error.
- Class 2: Technical task-specific evaluation: Quantifies technical performance of an algorithm on a clinical task using measures such as accuracy, repeatability, and reproducibility.
  Example: Evaluating accuracy on the task of lesion detection with the AI low-dose PET reconstructed images.
- Class 3: Clinical evaluation: Quantifies the algorithm's efficacy to assist in making clinical decisions. AI algorithms that claim improvements in making diagnostic, predictive, prognostic, or therapeutic decisions require clinical evaluation.
  Example: Evaluating the AI reconstruction algorithm on the task of clinically diagnosing patients referred with the suspicion of recurrence of cancer.
- Class 4: Postdeployment evaluation: Monitors algorithm performance in dynamic real-world settings after clinical deployment. This may also assess off-label use, such as the algorithm's utility in populations and diseases beyond the original claim or with improved imaging cameras and reconstructions that were not used during training. Additionally, this evaluation assesses clinical utility and value over time.
  Example: Evaluating whether the AI PET reconstruction algorithm remains effective over time after clinical deployment.

In the subsections below, for each class of evaluation, we provide the key objectives, the best practices for study design (including determining study type, data collection, defining a reference standard, and choosing FoMs (Fig. 4)), and finally, a generic structure for the claim.

III.1. Proof-of-Concept (POC) Evaluation

III.1.1. Objective: The objective of POC evaluation is to quantitatively demonstrate the technologic innovations of newly developed AI algorithms using task-agnostic FoMs and provide evidence that motivates clinical task-specific evaluation. Clinical or task-specific technical claims should not be put forth based on POC evaluation.

Rationale for Task-Agnostic Objective: A newly developed AI algorithm may be suitable for multiple clinical tasks. For example, a segmentation algorithm may be applicable to radiation therapy planning, estimating volumetric or radiomic features, or monitoring therapy response. Evaluating the algorithm on all these tasks would

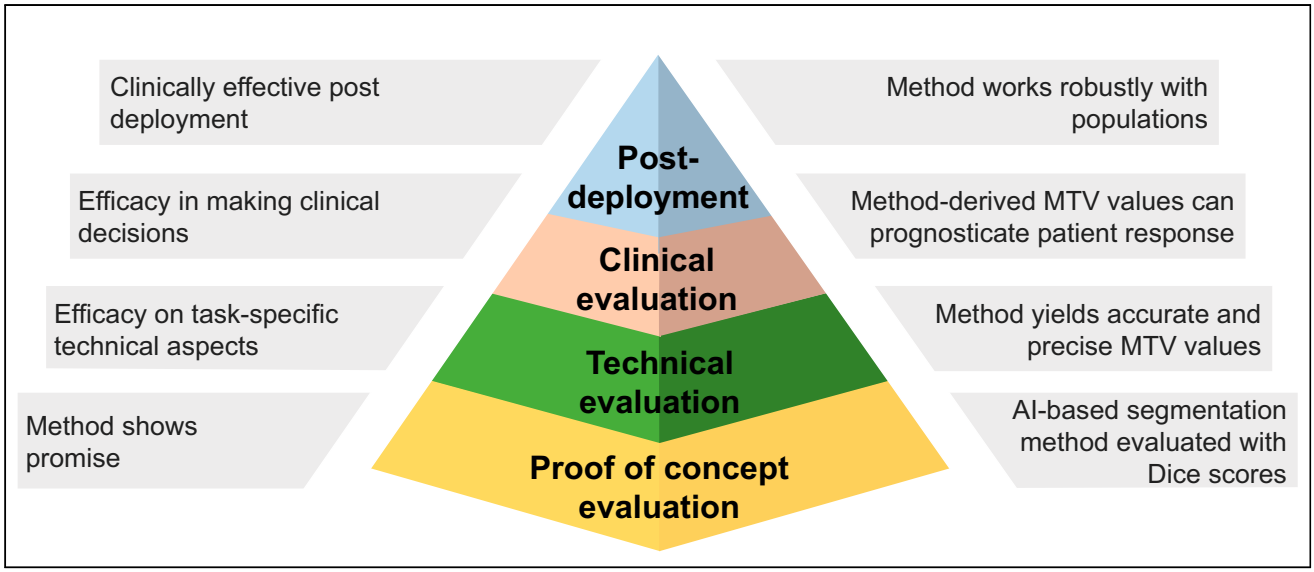

FIGURE 3. Framework for evaluation of AI-based algorithms. The left of the pyramid provides a brief description of the phase, and the right provides an example of evaluating an AI-based segmentation algorithm on the task of evaluating MTV using this framework.

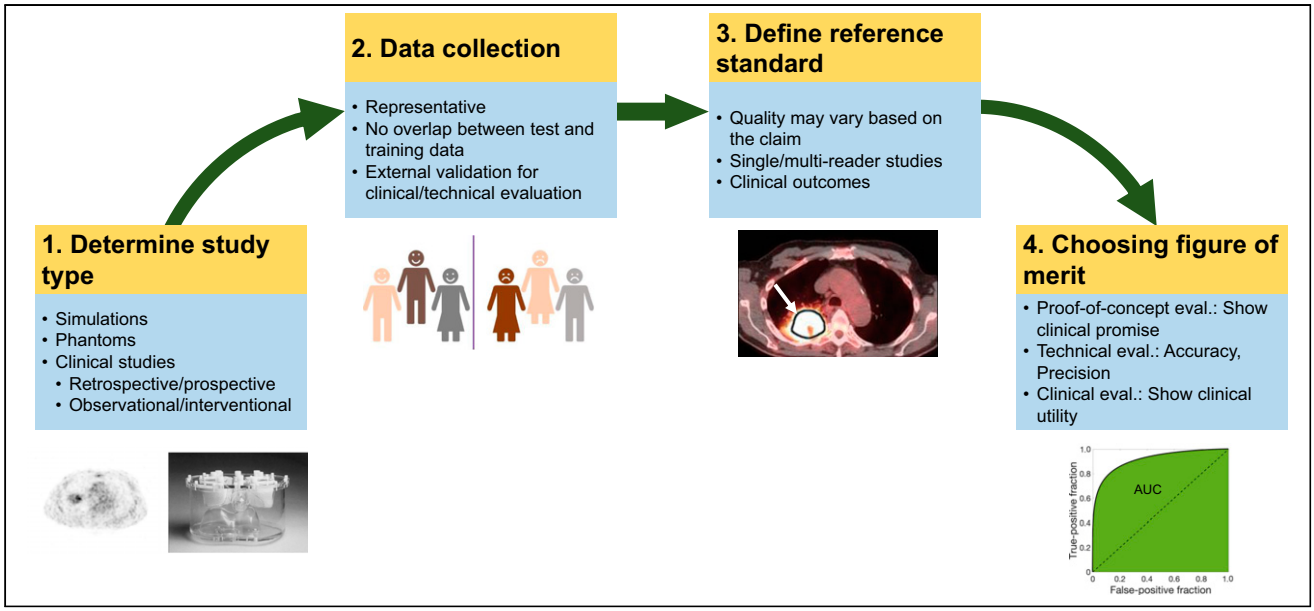

FIGURE 4. Elements of study design for each class of evaluation.

require multiple studies. Further, necessary resources (such as a large, representative dataset) may not be available to conduct these studies. Thus, a task-agnostic objective facilitates timely dissemination and widens the scope of newly developed AI methods.

III.1.2. Study Design: The following are recommended best practices to conduct POC evaluation of an AI algorithm. Best practices to develop the algorithm are covered in the companion paper (*16*).

Data Collection: In POC evaluation, the study can use realistic simulations, physical phantoms, or retrospective clinical or research data, usually collected for a different purpose, for example, routine diagnosis. The data used for evaluation may come from the development cohort, that is, the same overall cohort that the training and validation cohorts were drawn from. However, there must be no overlap between these data. Public databases, such as those available at The Cancer Imaging Archive (*18*) and from medical image analysis challenges, such as at https://grand-challenge.org, can also be used.

Defining Reference Standard: For POC evaluations conducted with simulation and physical phantoms, the ground truth is known. For clinical data, curation by readers may be used, but that may not be of the highest quality. For example, curations by a single reader may be sufficient.

Testing Procedure: The testing procedure should be designed to demonstrate promising technologic innovation. The algorithm should thus be compared against reference and/or standard-of-care methods and preferably other state-of-the-art algorithms.

Figures of Merit: While the evaluation is task-agnostic, the FoMs should be carefully chosen to show promise for progression to clinical task evaluation. For example, evaluating a new denoising algorithm that overly smooths the image at the cost of resolution using the FoM of contrast-to-noise ratio may be misleading. In those cases, a FoM such as structural similarity index may be more relevant. We recommend evaluation of the algorithms using multiple FoMs. A list of some FoMs is provided in Supplemental Table 1.

III.1.3. Output Claim of the POC Study: The claim should state the following:

- The application (e.g., segmentation, reconstruction) for which the method is proposed.
- The patient population.
- The imaging and image analysis protocol(s).
- Process to define reference standard.
- Performance as quantified with a task-agnostic evaluation metric.

We reemphasize that the POC study claim should not be interpreted as an indication of the algorithm's expected performance in a clinical setting or on any clinical task.

Example Claim: Consider the evaluation of a new segmentation algorithm. The claim could read as follows:

> "An AI-based PET tumor-segmentation algorithm evaluated on 50 patients with locally advanced breast cancer acquired on a single scanner with single-reader evaluation yielded mean Dice scores of 0.78 (95% CI 0.71-0.85)."

III.2 Technical Task-Specific Evaluation

III.2.1. Objective: The objective of technical task-specific evaluation is to evaluate the technical performance of an AI algorithm on specific clinically relevant tasks such as those of detection and quantification using FoMs that quantify aspects such as accuracy (discrimination accuracy for detection task and measurement bias for quantification task) and precision (reproducibility and repeatability). The objective is not to assess the utility of the method in clinical decision making, because clinical decision making is a combination of factors beyond technical aspects, such as prior clinical history, patient biology, other patient characteristics (age/sex/ethnicity), and results of other clinical tests. Thus, this evaluation does not consider clinical outcomes.

For example, evaluating the accuracy of an AI-based segmentation method to measure MTV would be a technical efficacy study. This study would not assess whether more accurate MTV measurement led to any change in clinical outcome.

III.2.2. Study Design: Given the goal of evaluating technical performance, the evaluation should be performed in controlled settings. Practices for designing such studies are outlined below. A framework and summary of tools to conduct these studies in the context of PET is provided in Jha et al. (*10*).

TABLE 1
Technical Evaluation: Comparison of Different Study Types, Associated Trade-Offs, and Criteria That Can Be Evaluated with the Study Type

		Simulation studies	Physical phantoms	Clinical studies
Advantage	Known ground truth	Y	Y	Rarely
	Scanner-based		Y	Y
	Model patient biology	Yes, but limited		Y
	Model population variability	Y		Y
Criterion that can be evaluated	Accuracy	Y	Y	
	Repeatability/reproducibility/noise sensitivity with multiple replicates	Y	Y	
	Repeatability/reproducibility/noise sensitivity with test–retest replicates		Y	Yes and recommended
	Biologic repeatability/reproducibility/noise sensitivity			Y
Other factors to consider	Costs	Low	Medium	High
	Time	Low	Medium	High
	Confidence about clinical realism	Low	Medium	High

Study Type: A technical evaluation study can be conducted through the following mechanisms:

1. Realistic simulations are studies conducted with anthropomorphic digital phantoms simulating patient populations, where measurements corresponding to these phantoms are generated using accurately simulated scanners. This includes virtual clinical trials, which can be used to obtain population-based inferences (*19–21*).
2. Anthropomorphic physical phantom studies are conducted on the scanners with devices that mimic the human anatomy and physiology.
3. Clinical-data-based studies where clinical data are used to evaluate the technical performance of an AI algorithm, for example, repeatability of an AI algorithm measuring MTV in test-retest PET scans.

The tradeoffs with these 3 study types are listed in Table 1. Each study type can be single or multiscanner/center, depending on the claim:

- Single-center/single-scanner studies are typically performed with a specific system, image acquisition, and reconstruction protocol. In these studies, the algorithm performance can be evaluated for variability in patients, including different demographics, habitus, or disease characteristics, while keeping the technical aspects of the imaging procedures constant. These studies can measure the sensitivity of the algorithm to patient characteristics. They can also study the repeatability of the AI algorithm. Reproducibility may be explored by varying factors such as reconstruction settings.
- Multicenter/multiscanner studies are mainly suitable to explore the sensitivity of the AI algorithm to acquisition variabilities, including variability in imaging procedures, systems, reconstruction methods and settings, and patient demographics if using clinical data. Typically, multicenter studies are performed to improve patient accrual in trials and therefore the same inclusion and exclusion criteria are applied to all centers. Further, multicenter studies can help assess the need for harmonization of imaging procedures and system performances.

Data Collection:

- Realistic simulation studies: To conduct realistic simulations, multiple digital anthropomorphic phantoms are available (*22*). In virtual clinical trial–based studies, the distribution of simulated image data should be similar to that observed in clinical populations. For this purpose, parameters derived directly from clinical data can be used during simulations (*4*). Expert reader-based studies can be used to validate realism of simulations (*23*).
  Next, to simulate the imaging systems, tools such as GATE (*24*), SIMIND (*25*), SimSET (*26*), PeneloPET (*27*), and others (*10*) can be used. Different system configurations, including those replicating multicenter settings, can be simulated. If the methods use reconstruction, then clinically used reconstruction protocols should be simulated. Simulation studies should not use data used for algorithm training/validation.
- Anthropomorphic physical phantom studies: For clinical relevance, the tracer uptake and acquisition parameters when imaging these phantoms should mimic that in clinical settings. To claim generalizable performance across different scanner protocols, different clinical acquisition and reconstruction protocols should be used. A phantom used during training should not be used during evaluation irrespective of changes in acquisition conditions between training and test phases.
- Clinical data: Technical evaluation studies will typically be retrospective. Use of external datasets, such as those from an institution or scanner not used for method training/validation, is

recommended. Public databases may also be used. Selection criteria should be defined.

Process to Extract Task-Specific Information:

- Classification task: Performance of AI-based reconstruction or postreconstruction algorithms should ideally be evaluated using psychophysic studies by expert readers. Methods such as 2 alternative forced-choice tests and ratings-scale approaches could be used. When human-observer studies are infeasible, validated numeric anthropomorphic observers, such as the channelized Hotelling observer with anthropomorphic channels, could be used (*11,28,29*). This may be a better choice than using untrained human observers, who may yield misleading measures of task performance. AI algorithms for optimizing instrumentation/acquisition can be evaluated directly on projection data. This provides the benefit that the evaluation would be agnostic to the choice of the reconstruction and analysis method (*30,31*). In this case, observers that are optimal in some sense, such as the ideal observer (which yields the maximum possible area under the receiver-operating-characteristics [ROC] curve [AUC] of all observers) should be used (*28*). The ideal observer can be challenging to compute in clinical settings, and to address this different strategies are being developed (*32,33*). An example of evaluating a hypothetical AI method for improving timing resolution in a time-of-flight PET system is presented in Jha et al. (*10*).
- Quantification task: The task should be performed using optimal quantification procedures to ensure that the algorithm evaluation is not biased due to a poor quantification process. Often, performing quantification requires an intermediate manual step. For example, the task of regional uptake quantification from reconstructed images may require manual delineation of regions of interest. Expert readers should perform these steps. Nuclear medicine images are noisy and corrupted by image-degrading processes. Thus, the process of quantification should account for the physics and statistical properties of the measured data. For example, if evaluating a segmentation algorithm on the task of quantifying a certain feature from the image, the process of estimating that feature should account for the image-degrading processes and noise (*10*). Maximum-likelihood estimation methods could be an excellent choice since they are often unbiased and if an efficient estimator exists, they are efficient (*11*). If using prior information on the parameters to be estimated, maximum-a-posteriori (*34*) and posterior-mean (*35*) estimators could be used. In several cases, measuring quantitative features directly from projection data may yield optimal quantification (*36,37*) and can be considered.
- Joint classification/quantification task: These tasks should again be performed optimally. If manual inputs are needed for the classification or quantification component of the task, these should be provided by expert readers. Numeric observers such as channelized scanning linear observers (*38*) and those based on deep learning (*39*) can also be used.

Defining a Reference Standard: For simulation studies, the ground-truth is known. Experimental errors may arise when obtaining ground truth from physical-phantom studies, and preferably, these should be modeled during the statistical analysis. For clinical studies, ground truth is commonly unavailable. A common workaround is to define a reference standard. The quality of curation to define this standard should be high. When the reference standard is expert defined, multireader studies are preferred where the readers have not participated in the training of the algorithm, and where each reader independently interprets images, blinded to the results of the AI algorithm and the other readers (*40*). In other cases, the reference standard may be the current clinical practice. Finally, another approach is to use no-gold-standard evaluation techniques, which have shown ability to evaluate algorithm performance on quantification tasks without ground truth (*41–43*).

Figures of Merit: A list of FoMs for different tasks is provided in Supplemental Table 2. Example FoMs include AUC to quantify accuracy on classification tasks, bias, variance, and ensemble mean square error to quantify accuracy, precision, and overall reliability on quantification tasks, and area under the estimation ROC curve for joint detection/classification tasks. Overall, we recommend the use of objective task-based measures to quantify performance, and not measures that are subjective and do not correspond to the clinical task. For a multicenter study, variability of these FoMs across centers, systems, or observers should be reported.

III.2.3. Output Claim from Evaluation Study: The claim will consist of the following components:

- The clinical task (detection/quantification/combination of both) for which the algorithm is evaluated.
- The study type (simulation/physical phantom/clinical).
- If applicable, the imaging and image analysis protocol.
- If clinical data, process to define ground truth.
- Performance, as quantified with task-specific FoMs.

Example Claim: Consider the same automated segmentation algorithm as mentioned in the proof-of-concept section being evaluated to estimate MTV. The claim could be:

> "An AI-based fully automated PET tumor-segmentation algorithm yielded MTV values with a normalized bias of X% (95% confidence intervals) as evaluated using physical-phantom studies with an anthropomorphic thoracic phantom conducted on a single scanner in a single center."

III.3 Clinical Evaluation

III.3.1. Objective: Evaluate the impact of the AI algorithm on making clinical decisions, including diagnostic, prognostic, predictive, and therapeutic decisions for primary endpoints such as improved accuracy or precision in measuring clinical outcome. While technical evaluation is geared toward quantifying the performance of a technique in controlled settings, clinical evaluation investigates clinical utility in a practical setting. This evaluation will assess the added value that the AI algorithm brings to clinical decision making.

III.3.2. Study Design:

Study Type: The following study types can be used:

- Retrospective study: A retrospective study uses existing data sources. In a blinded retrospective study, readers analyzing the study data are blinded to the relevant clinical outcome. Retrospective studies are the most common mechanism to evaluate AI algorithms. Advantages of these studies include low costs and quicker execution. These studies can provide considerations for designing prospective studies. With rare diseases, these may be the only viable mechanism for evaluation. However, these studies cannot conclusively demonstrate causality between the algorithm output and the clinical outcome. Also, these studies may be affected by different biases such as patient-selection bias.
- Prospective observational study: In this study, the consequential outcomes of interest occur after study commencement, but the

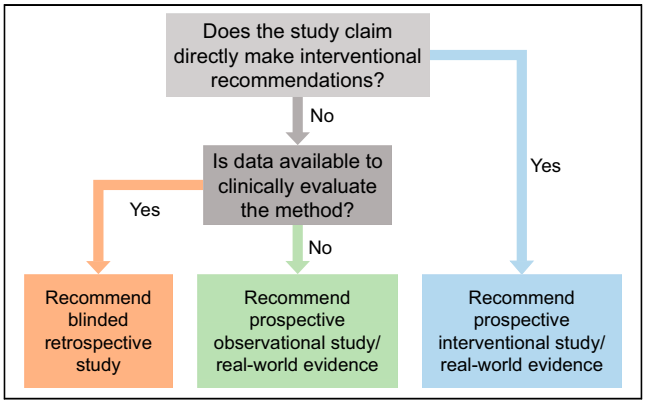

FIGURE 5. Flowchart to determine the clinical evaluation strategy.

decision to assign participants to an intervention is not influenced by the algorithm (*44*). These studies are often secondary objectives of a clinical trial.

- Prospective interventional study: In a prospective interventional study of an AI algorithm, the decision to assign the participant to an intervention depends on the AI algorithm output. These studies can provide stronger evidence for causation of the AI algorithm output to clinical outcome. The most common and strongest prospective interventional study design are randomized control trials, although other designs such as nonrandomized trials and quasiexperiments are possible (*45*). Randomized control trials are considered the gold standard of clinical evaluation but are typically logistically challenging, expensive, and time consuming, and should not be considered as the only means to ascertain and establish effective algorithms.
- Real-world postdeployment evaluation studies: These studies use real-world data from AI algorithms that have received regulatory clearance (*43*). Such studies have the potential to provide information on a wider patient population compared with a prospective interventional study. Moreover, the real-world data can be leveraged not only to improve performance of the initially cleared AI device but also to evaluate new clinical applications that require the same data or data similar to the initially cleared AI module, thus saving time and cost. The study design should be carefully crafted with a study protocol and analysis plan defined before retrieving/analyzing the real-world data (*46,47*), with special attention paid to negate bias (*48*).

Choosing the study type is a multifactorial decision (Fig. 5). To decide on the appropriate study type, we make a distinction between AI algorithms that make *direct* interventional recommendations (prescriptive AI) and those that do not (descriptive AI):

- A purely descriptive AI algorithm does not make direct interventional recommendations but may alter clinical decision making. The algorithms can be further categorized into those that describe the present (e.g., for diagnosis, staging, therapy response assessment) versus those that predict the future (e.g., prognosis of therapy outcome, disease progression, overall survival). There are close links between these 2 categories, and the line between them will likely be increasingly blurred in the era of AI: for example, more-refined AI-derived cancer staging that is trained with outcome data and therefore becomes highly predictive of outcome. A well-designed blinded retrospective study is sufficient to evaluate a purely descriptive AI system. However, if clinical data for a retrospective study do not exist, a prospective observational or real-world study is required.
- A prescriptive AI algorithm makes direct interventional recommendation(s). It may have no autonomy (i.e., only making a recommendation to a physician) or full autonomy (no supervision), or grades in between. For a prescriptive AI algorithm that is not autonomous, a prospective interventional study is recommended. A well-designed real-world study may be used as a substitute. However, for a fully autonomous prescriptive AI system of the future (e.g., fully automated therapy delivery), such a study may be required. Future studies and recommendations are needed for autonomous prescriptive AI systems, as the field is not mature enough. Thus, we limit the scope of this section to only those systems that have expert physician supervision.

Data Collection: An AI algorithm yielding strong performance using data from one institution may perform poorly on data from other institutions (*5*). Thus, we recommend that for clinical evaluation, test data should be collected from different, and preferably multiple, institutions. Results from external institutions can be compared with internal hold-out samples (data from the same institution not used for training) to evaluate generalizability. To avoid variation due to site selection used for the external validation, or random bias in internal sample selection, a leave-one-site repeated hold-out (e.g., 10-fold cross-validation) strategy can be used with a dataset that is completely independent from the training and validation dataset.

To demonstrate applicability over a certain target population, the collected data should be representative of that population in terms of demographics. When the goal is studying performance on a specific population subset (e.g., patients with large body mass indices) or checking sensitivity of the method to certain factors (e.g., patients with metallic implants), the other criteria for patient selection should be unbiased. This ensures that the evaluation specifically studies the effect of that factor.

In studies that are retrospective or based on real-world data, once a database has been set up corresponding to a target population using existing datasets, patients should be randomly selected from this database to avoid selection bias.

Sample-size considerations: The study must have a predefined statistical analysis plan (*49*). The sample size is task dependent. For example, if the claim of improved AUC with the AI method versus a non-AI approach or standard clinical analysis is studied, then the sample size will be dictated by the detection of the expected change between the 2 AUCs. Inputs required for power analysis to compute sample size may be obtained from POC and technical evaluation studies or separate pilot studies.

Defining a Reference Standard: For clinical evaluation, the reference standard should be carefully defined. This requires in-depth clinical and imaging knowledge of the data. Thus, medical experts should be involved in defining task-specific standards. Some reference standards are listed below:

- Clinical outcomes: Eventually the goal of imaging is to improve clinical outcomes. Outcomes such as overall survival, progression-free survival, major clinical events, and hospitalization could thus serve as gold standards, especially for demonstrating clinical utility in predictive and prognostic tasks. A decrease in the use of resources because of the AI tool with comparable outcomes could also be a relevant and improved outcome (e.g., fewer nonessential call back tests with AI).
- External standard: For disease diagnosis tasks, when available, an external standard such as invasive findings, for example, biopsy-pathology or invasive coronary angiography, or some

other definitive diagnosis (derived from other means than the images used) could be considered.

- Trained-reader-defined clinical diagnosis: For diagnostic tasks, expert reader(s) can be used to assess the presence/absence of the disease. Similar best practices as outlined for evaluating technical efficacy should be followed to design these studies. However, note that, unlike technical evaluation, here the goal is disease diagnosis. Thus, the readers should also be provided other factors that are used to make a clinical decision, such as the patient age, sex, ethnicity, other clinical factors that may impact disease diagnosis, and results from other clinical tests. Note that if the reference standard is defined using a standard-of-care clinical protocol, it may not be possible to claim improvement over this protocol. In such a case, agreement-based studies can be performed and concordance with these protocol results could be claimed within certain confidence limits. For example, to evaluate the ability of an AI-based transmission-less attenuation compensation algorithm for SPECT/PET, we may evaluate agreement of the estimates yielded by this algorithm with that obtained when a CT is used for attenuation compensation (*50*).

Figure of Merit: We recommend quantifying performance on strong, impactful, and objectively measurable endpoints such as improved accuracy or precision in measuring clinical outcome. The FoMs are summarized in Supplemental Table 2. To evaluate performance on diagnosis tasks, the FoMs of sensitivity, specificity, ROC curves, and AUC can be used. Since the goal is demonstrating the performance of the algorithm in clinical decision making, sensitivity and specificity may be clinically more relevant than AUC. To demonstrate clinical utility in predictive and prognostic decision making, in addition to AUC, FoMs that quantify performance in predicting future events such as Kaplan–Meier estimators, prediction risk score, and median time of future events can be used.

III.3.3. Output Claim from Clinical Evaluation Study: The claim will state the following:

- The clinical task for which the algorithm is evaluated.
- The patient population over which the algorithm was evaluated.
- The specific imaging and image-analysis protocol(s) or standards followed.
- Brief description of study design: Blinded/nonblinded, randomized/nonrandomized, retrospective/prospective/postdeployment, observational/interventional, number of readers.

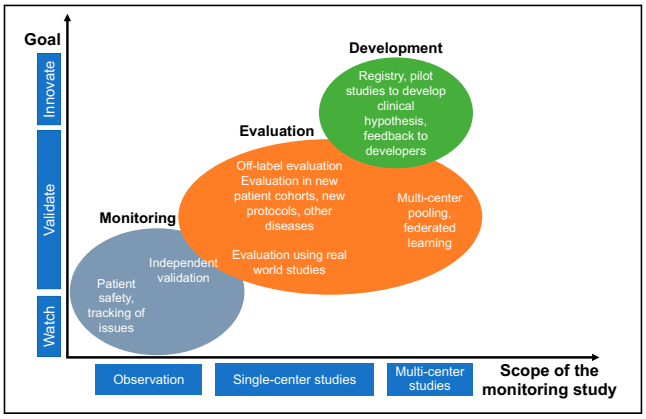

FIGURE 6. Chart showing the different objectives of postdeployment monitoring, grouped as a function of the scope and goal of the study.

- Process to define reference standard and FoM to quantify performance in clinical decision making.

Example Claims:

i. Retrospective study: The average AUC of 3 experienced readers on the task of detecting obstructive coronary artery disease from myocardial perfusion PET scans improved from X to Y, representing an estimated difference of Δ (95% CI for Δ), when using an AI-based diagnosis tool compared with not using this tool, as evaluated using a blinded retrospective study.

ii. Prospective observational study: Early change in MTV measured from FDG PET using an AI-based segmentation algorithm yielded an increase in AUC from X to Y, representing an estimated difference of Δ (95% CI for Δ) in predicting pathologic complete response in patients with stage II/III breast cancer, as evaluated using a nonrandomized prospective observational study.

iii. Prospective interventional study: Changes in PET-derived quantitative features estimated with the help of an AI algorithm during the interim stage of therapy were used to guide treatment decisions in patients with stage III non–small cell lung cancer. This led to an X% increase (95% CI) in responders than when the AI algorithm was not used to guide treatment decisions, as evaluated using a randomized prospective interventional study.

III.4. Postdeployment Evaluation

III.4.1. Objective: Postdeployment evaluation has multiple objectives. A key objective is monitoring algorithm performance after clinical deployment including evaluating clinical utility and value over time. Other objectives include off-label evaluation and collecting feedback for proactive development (Fig. 6).

III.4.2. Evaluation Strategies:

Monitoring: Quality and patient safety are critical factors in postdeployment monitoring of an AI algorithm. It is imperative to monitor devices and follow reporting guidelines (such as adverse events), recalls, and corrective actions. Fortunately, applicable laws and regulations require efficient processes in place. Often, logging is used to identify root causes for equipment failure. However, the concept of logging can be expanded: advanced logging mechanisms could be used to better understand use of an AI algorithm. A simple use case is logging the frequency of using an AI algorithm in clinical workflow. Measuring manual intervention for a workflow step that was designed for automation could provide a first impression of the performance in a clinical environment. However, more complex use cases may include the aggregation of data on AI algorithm performance and its impact on patient and disease management. For wider monitoring, feedback should be sought from customers, including focus groups, customer complaint and inquiry tracking, and ongoing technical performance benchmarking (*51*). This approach may provide additional evidence on algorithm performance and could assist in finding areas of improvements, clinical needs not well served or even deriving a hypothesis for further development. Advanced data logging and sharing must be compliant with applicable patient privacy and data protection laws and regulations.

Routinely conducted image-quality phantom studies also provide a mechanism for postdeployment evaluation by serving as sanity checks to ensure that the AI algorithm was not affected by a maintenance operation such as a software update. These studies could include assessing contrast or SUV recovery, absence of nonuniformities or artifacts, cold-spot recovery, and other specialized tests depending on the AI algorithm. Also, tests can be conducted to

TABLE 2
RELAINCE Guidelines

Class of evaluation	Recommendation
Proof of concept evaluation	Ensure no overlap between development and testing cohort.
	Check that ground-truth quality is reasonable.
	Provide comparison with conventional and state-of-the-art methods.
	Choose figures of merit that motivate further clinical evaluation.
Technical task-specific evaluation	Choose clinically relevant tasks: Detection/quantification/combination of both.
	Determine the right study type: Simulation/phantom/clinical.
	Ensure that simulation studies are realistic and account for population variability.
	Testing cohort should be external.
	Reference standard should be high quality and correspond to the task.
	Use a reliable strategy to extract task-specific information.
	Choose figures of merit that quantify task performance.
Clinical evaluation	Determine study type: Retrospective, prospective observational, prospective interventional, or postdeployment real-world studies.
	Testing cohort must be external.
	Collected data should represent the target population as stated in the claim.
	Reference standard should be high quality and be representative of those used for clinical decision making.
	Figure of merit should reflect performance on clinical decision making.
Postdeployment evaluation	Monitor devices and follow reporting guidelines.
	Consider phantom studies as sanity checks to assess routine performance.
	Periodically monitor data drift.
	For off-label evaluation, follow recommendations as in clinical/technical evaluation depending on objective.

ensure that there is a minimal or harmonized image quality as required by the AI tool for the configurations as stated in the claim.

AI systems likely will operate on data generated in nonstationary environments with shifting patient populations and clinical and operational practices changing over time (*9*). Postdeployment studies can help identify these dataset shifts and assess if recalibration or retraining of the AI method may be necessary to maintain performance (*52,53*). Monitoring the distribution of various patient population descriptors, including demographics and disease prevalence, can provide cues for detecting dataset shifts. In the case of changes in these descriptors, the output of the AI algorithm can be verified by physicians for randomly selected test cases. A possible solution to data shift is continuous learning of the AI method (*54*). In Supplemental Section B, we discuss strategies (*55–57*) to evaluate continuous-learning-based methods.

Off-Label Evaluation: Typically, an AI algorithm is trained and tested using a well-defined cohort of patients, in terms of patient demographics, applicable guidelines, practice preferences, reader expertise, imaging instrumentation, and acquisition and analysis protocols. However, the design of the algorithm may suggest acceptable performance in cohorts outside the intended scope of the algorithm. Here, a series of cases is appropriate to collect preliminary data that may suggest a more thorough trial. An example is a study where an AI algorithm that was trained on patients with lymphoma and lung cancer (*58*) showed reliable performance in patients with breast cancer (*59*).

Collecting Feedback for Proactive Development: Medical products typically have a long lifetime. This motivates proactive development and maintenance to ensure that a product represents state of the art throughout its lifetime. This may be imperative for AI, where technologic innovations are expected to evolve at a fast pace in the coming years. A deployed AI algorithm offers the opportunity to pool data from several users. Specifically, registry approaches enable cost-efficient pooling of uniform data, multicenter observational studies, and POC studies that can be used to develop a new clinical hypothesis or evaluate specific outcomes for particular diseases.

Figures of Merit: We provide the FoMs for the studies where quantitative metrics of success are defined.

- Monitoring study with clinical data: Frequency of clinical usage of the AI algorithm, number of times the AI-based method changed clinical decisions or affected patient management.
- Monitoring study with routine physical phantom studies: Since these are mostly sanity checks, FoMs similar to those used when evaluating POC studies may be considered. In case task-based evaluation is required, FoMs as provided in Supplemental Table 1 may be used.
- Off-label evaluation: FoMs similar to those used when performing technical and clinical evaluation may be considered.

IV. DISCUSSION

The key recommendations from this article are summarized in Table 2. These are referred to as the RELAINCE (Recommendations

for EvaLuation of AI for NuClear medicinE) guidelines, with the goal of improving the reliance of AI for clinical applications. Unlike other guidelines for the use of AI in radiology (*60–62*), these guidelines are exclusively focused on best practices for AI algorithm evaluation.

This report advocates that an evaluation study should be geared toward putting forth a claim. The objective of the claim can be guided by factors such as the degree of impact on patient management, level of autonomy, and the risk that the method poses to patients. Risk categories have been proposed for medical software by the International Medical Device Regulators Forum and subsequently adopted by the Food and Drug Administration (*63*). The proposed risk categories range from 1 (low risk) to 4 (highest risk) depending on the vulnerability of the patient and the degree of control that the software has in patient management. The pathway that a developing technology will take to reach clinical adoption will ultimately depend on which risk category it belongs to, and investigators should assess risk early during algorithm development and plan accordingly (*64*).

In this report, we have proposed a 4-class framework for evaluation. For clinical adoption, an algorithm may not need to pass through all classes. The POC evaluation is optional as the objective of this class is to only demonstrate promise for further evaluation. Further, not all these classes may be fully relevant to all algorithms. For example, an AI segmentation algorithm may require technical but not necessarily clinical evaluation for clinical adoption. The types of studies required for an algorithm will depend on the claim. For example, an AI algorithm that claims to make improvement in making clinical decisions will require clinical evaluation. For clinical acceptability of an AI algorithm, evaluating performance on clinical tasks is most important. POC, technical, and clinical evaluation could all be reported in the same multipart study.

The evaluation studies should preferably be multidisciplinary and include computational imaging scientists, physicians, physicists, and statisticians right from the study conception stage. Physicians should be closely involved because they are the end users of these algorithms. Previous publications have outlined the important role of physicians in evaluation of AI algorithms (*65*), including for task-based evaluation of AI algorithms for nuclear medicine (*10*).

The proposed best practices are generally applicable to evaluating a wide class of AI algorithms, including supervised, unsupervised, and semisupervised approaches. For example, we recommend that for even semisupervised and unsupervised learning algorithms, the algorithm should be evaluated on previously unseen data. Additionally, these best practices are broadly applicable to other machine learning as well as physics-based algorithms for nuclear medicine imaging. Further, whereas these guidelines are being proposed in the context of nuclear medicine imaging, they are also broadly applicable to other medical imaging modalities.

In addition to the above practices, we also recommend that in each class of evaluation, evaluation studies should attempt to assess the interpretability of the algorithm. In fact, rigorous evaluations may provide a mechanism to make the algorithm more interpretable. For example, a technical efficacy study may observe suboptimal performance of an AI-based denoising algorithm on the tumor-detection task. Then, the evaluation study could investigate the performance of the algorithm for different tumor properties (size/tumor-to-background ratio) on the detection task (*66*). This will provide insights on the working principles of the algorithm, thus improving the interpretability of the algorithm.

In summary, AI-based algorithms present an exciting toolset for advancing nuclear medicine. We envision that following these best practices for evaluation will assess suitability and provide confidence for clinical translation of these algorithms, and provide trust for clinical application, ultimately leading to improvements in the quality of health care.

DISCLOSURE

Sven Zuehlsdorff is a full-time employee of Siemens Medical Solutions USA, Inc. Nancy Obuchowski is a contracted statistician for QIBA. Tyler Bradshaw receives research support from GE Healthcare. Ronald Boellaard is (unpaid) scientific advisor for the EARL PET/CT accreditation program. Piotr Slomka has a research grant from Siemens Medical Solutions, is a consultant for IBA, and receives royalties from Cedars-Sinai for nuclear cardiology software. No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

The Task Force members thank Kyle J. Myers, PhD, for helpful discussions and Bonnie Clarke for all her help throughout this project.

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44–56.
2. Yang J, Sohn JH, Behr SC, Gullberg GT, Seo Y. CT-less direct correction of attenuation and scatter in the image space using deep learning for whole-body FDG PET: potential benefits and pitfalls. *Radiol Artif Intell.* 2020;3:e200137.
3. Yu Z, Rahman MA, Schindler T, et al. AI-based methods for nuclear-medicine imaging: need for objective task-specific evaluation [abstract]. *J Nucl Med.* 2020; 61(suppl 1):575.
4. Leung KH, Marashdeh W, Wray R, et al. A physics-guided modular deep-learning based automated framework for tumor segmentation in PET. *Phys Med Biol.* 2020; 65:245032.
5. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* 2018;15:e1002683.
6. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* 2018;178:1544–1547.
7. Noor P. Can we trust AI not to further embed racial bias and prejudice? *BMJ.* 2020;368:m363.
8. Reuzé S, Orlhac F, Chargari C, et al. Prediction of cervical cancer recurrence using textural features extracted from ^{18}F-FDG PET images acquired with different scanners. *Oncotarget.* 2017;8:43169–43179.
9. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med.* 2021;385:283–286.
10. Jha AK, Myers KJ, Obuchowski NA, et al. Objective task-based evaluation of artificial intelligence-based medical imaging methods: framework, strategies and role of the physician. *PET Clin.* 2021;16:493–511.
11. Barrett HH, Myers KJ. *Foundations of Image Science.* First vol. Wiley; 2004.
12. Liu Z, Mhlanga J, Siegel S, Jha A. Need for objective task-based evaluation of segmentation methods in oncological PET: a study with ACRIN 6668/RTOG 0235 multi-center clinical trial data [abstract]. *J Nucl Med.* 2022;63(suppl 2):2413.
13. KC P., Zeng R, Farhangi MM, Myers KJ. Deep neural networks-based denoising models for CT imaging and their efficacy. *Proc SPIE Med Imag.* 2021; 11595:105–117.
14. Myers KJ, Barrett HH, Borgstrom MC, Patton DD, Seeley GW. Effect of noise correlation on detectability of disk signals in medical imaging. *J Opt Soc Am A.* 1985;2:1752–1759.
15. Harris JL. Resolving power and decision theory*†. *J Opt Soc Am.* 1964;54:606–611.
16. Bradshaw TJ, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med.* 2021;63:500–510.
17. Garcia EV. SPECT attenuation correction: an essential tool to realize nuclear cardiology's manifest destiny. *J Nucl Cardiol.* 2007;14:16–24.

18. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* 2013;26: 1045–1057.
19. Abadi E, Segars W, Tsui BM, et al. Virtual clinical trials in medical imaging: a review. *J Med Imaging (Bellingham).* 2020;7:042805.
20. Yu Z, Rahman MA, Laforest R, Norris SA, Jha AK. A physics and learning-based transmission-less attenuation compensation method for SPECT. *Proc SPIE Med Imag.* 2021;11595:1159512.
21. Badano A, Graff CG, Badal A, et al. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Netw Open.* 2018;1:e185474.
22. Kainz W, Neufeld E, Bolch WE, et al. Advances in computational human phantoms and their applications in biomedical engineering: a topical review. *IEEE Trans Radiat Plasma Med Sci.* 2019;3:1–23.
23. Liu Z, Laforest R, Moon H, et al. Observer study-based evaluation of a stochastic and physics-based method to generate oncological PET images. *Proc SPIE Med Imag.* 2021;11599:1159905.
24. Jan S, Santin G, Strul D, et al. GATE: a simulation toolkit for PET and SPECT. *Phys Med Biol.* 2004;49:4543–4561.
25. Ljungberg M, Strand S, King M. The SIMIND Monte Carlo program. In: Ljungberg M, Strand S-E, King MA, eds. *Monte Carlo Calculation in Nuclear Medicine: Applications in Diagnostic Imaging.* CRC Press: 1998:145–163.
26. Lewellen T, Harrison R, Vannoy S. The SimSET program. In: Ljungberg M, Strand S-E, King MA, eds. *Monte Carlo Calculations in Nuclear Medicine: Applications in Diagnostic Imaging.* Vol. 87. CRC Press: 2012.
27. España S, Herraiz JL, Vicente E, Vaquero JJ, Desco M, Udias JM. PeneloPET, a Monte Carlo PET simulation tool based on PENELOPE: features and validation. *Phys Med Biol.* 2009;54:1723–1742.
28. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. *Proc Natl Acad Sci USA.* 1993;90:9758–9765.
29. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J Opt Soc Am A Opt Image Sci Vis.* 2001;18:473–488.
30. Gross K, Kupinski M, Peterson T, Clarkson E. *Optimizing a Multiple-Pinhole SPECT System Using the Ideal Observer.* Vol. 5034. SPIE; 2003.
31. Rong X, Ghaly M, Frey EC. Optimization of energy window for ^{90}Y bremsstrahlung SPECT imaging for detection tasks using the ideal observer with model-mismatch. *Med Phys.* 2013;40:062502.
32. Clarkson E, Shen F. Fisher information and surrogate figures of merit for the task-based assessment of image quality. *J Opt Soc Am A Opt Image Sci Vis.* 2010;27: 2313–2326.
33. Li X, Jha AK, Ghaly M, Link JM, Frey E. Use of sub-ensembles and multi-template observers to evaluate detection task performance for data that are not multivariate normal. *IEEE Trans Med Imaging.* 2017;36:917–929.
34. Whitaker MK, Clarkson E, Barrett HH. Estimating random signal parameters from noisy images with nuisance parameters: linear and scanning-linear methods. *Opt Express.* 2008;16:8150–8173.
35. Liu Z, Mhlanga JC, Laforest R, Derenoncourt P-R, Siegel BA, Jha AK. A Bayesian approach to tissue-fraction estimation for oncological PET segmentation. *Phys Med Biol.* 2021;66:10.1088/1361-6560/ac01f4.
36. Carson RE. A maximum likelihood method for region-of-interest evaluation in emission tomography. *J Comput Assist Tomogr.* 1986;10:654–663.
37. Li Z, Benabdallah N, Abou D, et al. A projection-domain low-count quantitative SPECT method for alpha-particle emitting radiopharmaceutical therapy. arxiv, Cornell University, website. https://arxiv.org/abs/2107.00740. Revised May 11, 2022. Accessed August 3, 2022.
38. Tseng H-W, Fan J, Kupinski MA. Combination of detection and estimation tasks using channelized scanning linear observer for CT imaging systems. *Proc SPIE Med Imag.* 2015;9416:94160H.
39. Li K, Zhou W, Li H, Anastasio MA. A Hybrid approach for approximating the ideal observer for joint signal detection and estimation tasks by use of supervised learning and markov-chain monte carlo methods. *IEEE Trans Med Imaging.* 2022;41:1114–1124.
40. Miller DP, O'shaughnessy KF, Wood SA, Castellino RA. Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions. *Proc SPIE Med Imag.* 2004;5372:173–184.
41. Hoppin JW, Kupinski MA, Kastis GA, Clarkson E, Barrett HH. Objective comparison of quantitative imaging modalities without the use of a gold standard. *IEEE Trans Med Imaging.* 2002;21:441–449.
42. Jha AK, Caffo B, Frey EC. A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods. *Phys Med Biol.* 2016; 61:2780–2800.
43. Jha AK, Mena E, Caffo B, et al. Practical no-gold-standard evaluation framework for quantitative imaging methods: application to lesion segmentation in positron emission tomography. *J Med Imaging (Bellingham).* 2017;4:011011.
44. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand S-L. Prospective observational studies to assess comparative effectiveness: The ISPOR good research practices task force report. *Value Health.* 2012;15:217–230.
45. Thiese MS. Observational and interventional study design types; an overview. *Biochem Med (Zagreb).* 2014;24:199–210.
46. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence: what is it and what can it tell us? *N Engl J Med.* 2016;375:2293–2297.
47. US Food Drug Administration. *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices.* 2017. FDA website. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices. Accessed August 19, 2022.
48. Tarricone R, Boscolo PR, Armeni P. What type of clinical evidence is needed to assess medical devices? *Eur Respir Rev.* 2016;25:259.
49. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ.* 2009;339:b4184.
50. Shi L, Onofrey JA, Liu H, Liu YH, Liu C. Deep learning-based attenuation map generation for myocardial perfusion SPECT. *Eur J Nucl Med Mol Imaging.* 2020; 47:2383–2395.
51. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: Summary and recommendations. *J Am Coll Radiol.* 2021;18:413–424.
52. Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc.* 2019;26:1448–1457.
53. Feng J. Learning to safely approve updates to machine learning algorithms. *Proc Conf on Health, Inference, and Learning.* 2021:164–173.
54. Baweja C, Glocker B, Kamnitsas K. Towards continual learning in medical imaging. arxiv, Cornell University, website. https://arxiv.org/abs/1811.02496. Submitted November 26, 2018. Accessed August 3, 2022.
55. Díaz-Rodríguez N, Lomonaco V, Filliat D, Maltoni D. Don't forget, there is more than forgetting: new metrics for continual learning. arxiv, Cornell University, website. https://arxiv.org/abs/1810.13166. Submitted October 31, 2018. Accessed August 3, 2022.
56. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. arxiv, Cornell University, website. https://arxiv.org/abs/1312.6211v3. Revised March 4, 2015. Accessed August 3, 2022.
57. Chaudhry A, Dokania PK, Ajanthan T, Torr PH. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *ECCV.* 2018:532–547.
58. Sibille L, Seifert R, Avramovic N, et al. ^{18}F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology.* 2020;294:445–452.
59. Weber M, Kersting D, Umutlu L, et al. Just another "Clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer. *Eur J Nucl Med Mol Imaging.* 2021;48:3141–3150.
60. Dikici E, Bigelow M, Prevedello LM, White RD, Erdal BS. Integrating AI into radiology workflow: levels of research, production, and feedback maturity. *J Med Imaging (Bellingham).* 2020;7:016502.
61. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell.* 2020;2:e200029.
62. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol.* 2021; 31:3786–3796.
63. *Software as a Medical Device (SaMD): Clinical Evaluation.* Center for Devices and Radiological Health, United States Food and Drug Administration; 2017. FDA website. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation. Accessed August 19, 2022.
64. *Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and de Novo Classifications: Guidance for Industry and Food and Drug Administration Staff.* Center for Devices and Radiological Health, USA Food and Drug Administration; 2012. FDA website. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/factors-consider-when-making-benefit-risk-determinations-medical-device-premarket-approval-and-de. Accessed August 19, 2022.
65. Rubin DL. Artificial intelligence in imaging: The radiologist's role. *J Am Coll Radiol.* 2019;16:1309–1317.
66. Yu Z, Rahman MA, Jha AK. Investigating the limited performance of a deep-learning-based SPECT denoising approach: an observer study-based characterization. *Proc SPIE Med Imag.* 2022. 12035:120350D.