Journal of Nuclear Medicine, published on May 26, 2022 as doi:10.2967/jnumed.121.263239

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], Ronald Boellaard[15]

[1]Department of Biomedical Engineering and Mallinckrodt Institute of Radiology, Washington University in St. Louis, USA
[2]Department of Radiology, University of Wisconsin-Madison, USA
[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, USA
[6]Department of Radiology and Biomedical Imaging, Yale University, USA
[7]Quantitative Health Sciences, Cleveland Clinic, Cleveland, USA
[8]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, USA
[9]Department of Imaging, Medicine, and Cardiology, Cedars-Sinai Medical Center, USA
[10]Departments of Radiology and Physics, University of Iowa, USA
[11]Mallinckrodt Institute of Radiology, Washington University in St. Louis, USA
[12]Department of Biomedical Engineering, Washington University in St. Louis, USA
[13]Siemens Medical Solutions USA, Inc., Hoffman Estates, USA
[14]Departments of Radiology and Physics, University of British Columbia, Canada
[15]Department of Radiology & Nuclear Medicine, Cancer Centre Amsterdam, Amsterdam University Medical Centers, Netherlands

Running title: Best practices AI Evaluation

Corresponding author:
Abhinav K. Jha, PhD
Assistant Professor of Biomedical Engineering and of Radiology
Department of Biomedical Engineering
Mallinckrodt Institute of Radiology
Washington University in St. Louis
Email: a.jha@wustl.edu
Phone: 314-273-2655

Word count ~ 9600 words

Financial support: None (individual COIs for each author are listed in Disclosures)

Keywords: Artificial intelligence, Evaluation, Best practices, clinical task, PET, SPECT, technical efficacy, clinical decision making, post-deployment, generalizability.

Noteworthy

1. AI algorithms should be evaluated on clinical tasks.
2. 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.
3. We propose a four-class framework that evaluates AI algorithms for nuclear-medicine imaging on clinical tasks and yields a claim. The four classes in the framework include promise, technical, clinical, and post-deployment evaluation of AI algorithms.
4. We provide best practices for determining study type, data collection, defining reference standard, and choosing figures of merit for each class of evaluation.
5. Key recommendations are summarized as the RELAINCE (Recommendations for EvaLuation of AI for NuClear medicinE) guidelines.

ABSTRACT

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 four-class framework to evaluate AI algorithms for promise, technical task-specific efficacy, clinical decision making, and post-deployment 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 taskforce Evaluation team, which consisted of nuclear-medicine physicians, physicists, computational imaging scientists, and representatives from industry and regulatory agencies.

INTRODUCTION

Artificial intelligence (AI)-based algorithms are showing tremendous promise across multiple aspects of nuclear medicine, including image acquisition, reconstruction, post-processing, 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 post-processing 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 typically not explicit and thus not easily interpretable, leading to unpredictability in output. This leads to multiple unique challenges. First, AI algorithms may yield results that may 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*). Such malfunctioning can adversely impact clinical utility. 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 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 post-deployment 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.

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 taskforce. 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, post-processing, 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 three categories: classification, quantification, or a combination of both (*10,11*). An oncological PET image may be acquired for the task of tumor-stage classification or for quantification of tracer uptake in tumor. However, current AI-algorithm evaluation strategies are often task agnostic. For example, AI algorithms for reconstruction and post-processing 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 to 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
- Figure of merit 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 four classes: proof-of-concept, technical, clinical and post-deployment 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*).

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 Fig. 1 and described below.

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 three 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 oncological 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 numerical or statistical feature is estimated from the patient image. Examples include quantifying standardized uptake value, metabolic tumor volume, intra-lesion heterogeneity or kinetic parameters from oncological PET images.

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 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 biological variables.

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 back-projection 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 these parameters. Further, if the algorithm was evaluated across multiple scanners, or with multiple protocols, that should be specified.

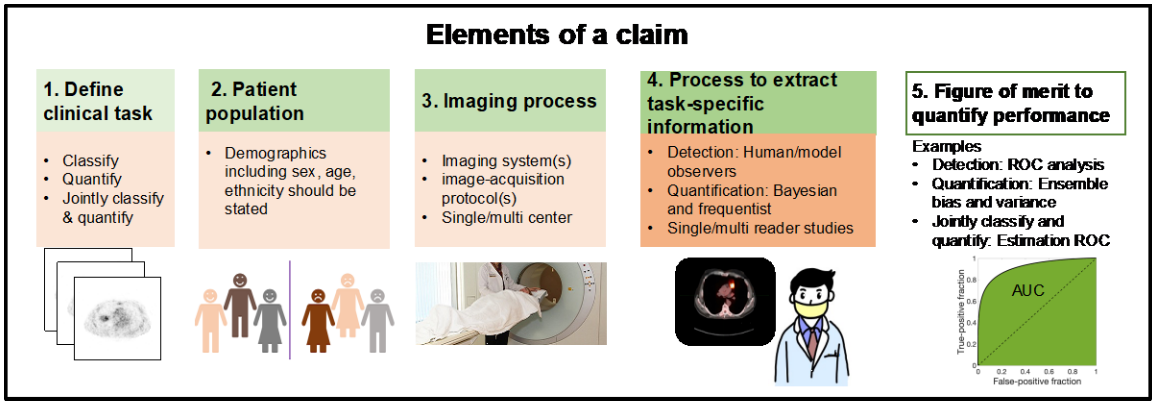

Elements of a claim
1. Define clinical task
• Classify
• Quantify
• Jointly classify & quantify
2. Patient population
• Demographics including sex, age, ethnicity should be stated
3. Imaging process
• Imaging system(s)
• image-acquisition protocol(s)
• Single/multi center
4. Process to extract task-specific information
• Detection: Human/model observers
• Quantification: Bayesian and frequentist
• Single/multi reader studies
5. Figure of merit to quantify performance
Examples
• Detection: ROC analysis
• Quantification: Ensemble bias and variance
• Jointly classify and quantify: Estimation ROC
AUC
True-positive fraction
False-positive fraction
0
0.2
0.4
0.6
0.8
1

Fig. 1: The components of a claim

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.

Figure of Merit (FoM) to Quantify Task Performance

FoMs quantitatively describe the algorithms 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 algorithms 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 multi-center, define the image acquisition and analysis protocols used, as well as the competency of the observer performing the evaluation study. Fig. 2 presents a schematic showing how different kinds of generalizability could be established. Some key points from this figure are listed below:

- 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.

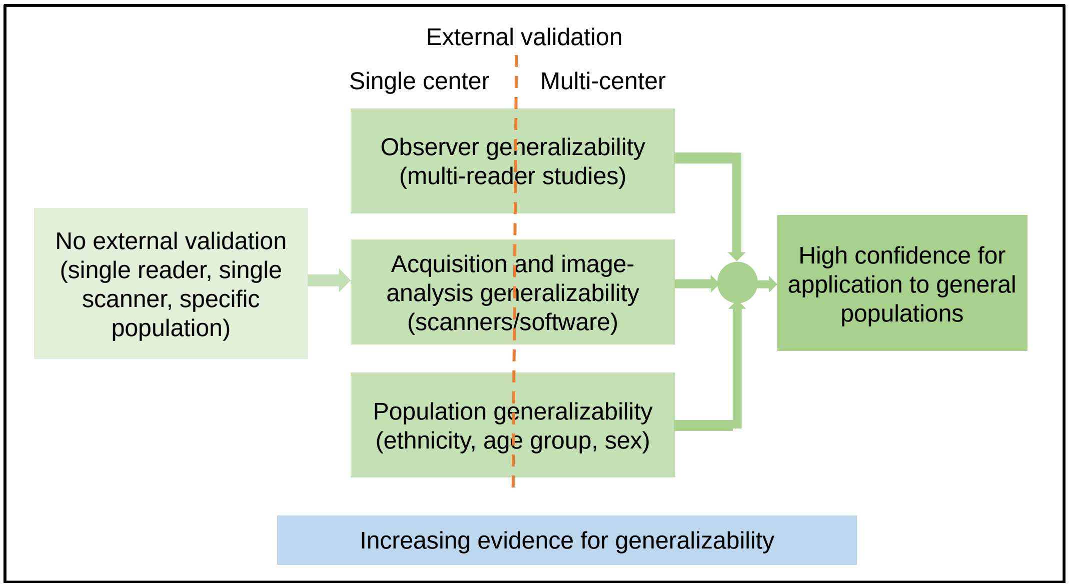

External validation
Single center
Multi-center
Observer generalizability (multi-reader studies)
No external validation (single reader, single scanner, specific population)
Acquisition and image-analysis generalizability (scanners/software)
High confidence for application to general populations
Population generalizability (ethnicity, age group, sex)
Increasing evidence for generalizability

Fig. 2: increasing levels of rigor of evaluation, and how they in turn provide increased confidence in the generalizability

- 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.
- Multi-center studies provide higher confidence about generalizability compared to single-center studies since they typically include some level of external validation (patients from different geographical locations/different scanners/different readers).

METHODS FOR EVALUATION

The evaluation framework for AI algorithms is provided in Fig. 3. The four 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. Fig. 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 (Supplemental Fig. 1) (*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: Post-deployment evaluation: Monitor 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.

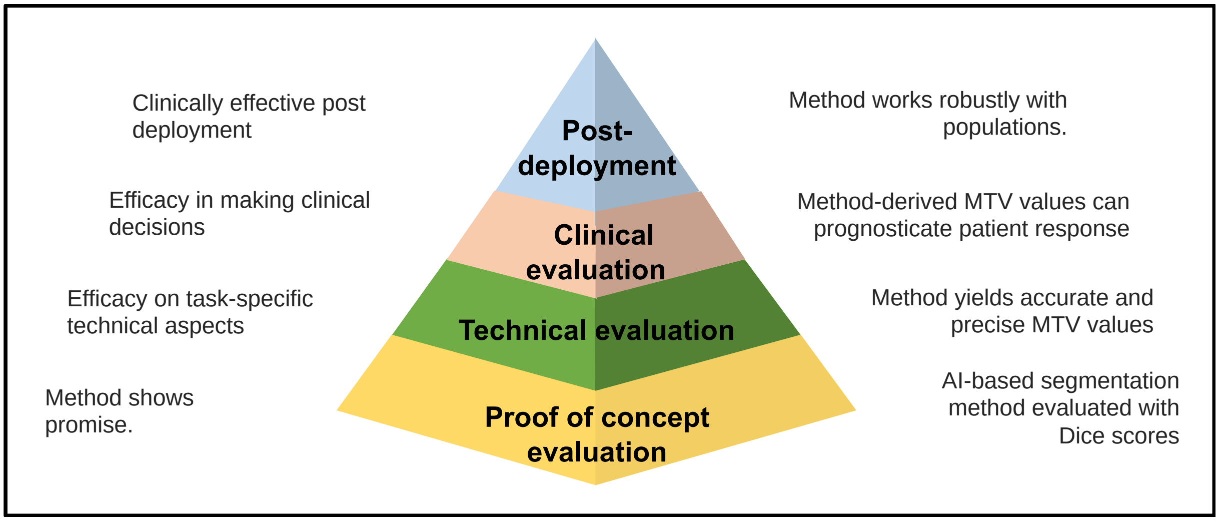

Clinically effective post deployment
Post-deployment
Method works robustly with populations.
Efficacy in making clinical decisions
Clinical evaluation
Method-derived MTV values can prognosticate patient response
Efficacy on task-specific technical aspects
Technical evaluation
Method yields accurate and precise MTV values
Method shows promise.
Proof of concept evaluation
AI-based segmentation method evaluated with Dice scores

Fig. 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 metabolic tumor volume (MTV) using this framework.

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.

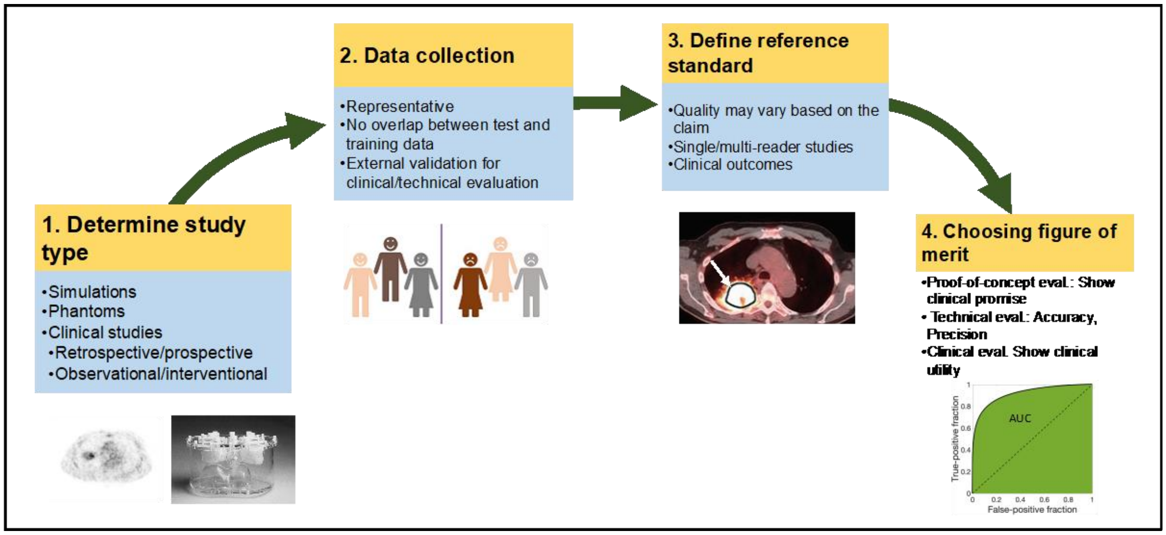

1. Determine study type
•Simulations
•Phantoms
•Clinical studies
•Retrospective/prospective
•Observational/interventional
2. Data collection
•Representative
•No overlap between test and training data
•External validation for clinical/technical evaluation
3. Define reference standard
•Quality may vary based on the claim
•Single/multi-reader studies
•Clinical outcomes
4. Choosing figure of merit
•Proof-of-concept eval.: Show clinical promise
• Technical eval.: Accuracy, Precision
•Clinical eval. Show clinical utility
AUC
True-positive fraction
False-positive fraction
0
0.2
0.4
0.6
0.8
1

Fig. 4: Elements of study design for each class of evaluation

M.1 Proof-of-concept (POC) evaluation

Objective

Quantitatively demonstrate the technological 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 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.

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, and/or retrospective clinical or research data, usually collected for a different purpose, e.g., routine diagnosis. The data used for evaluation may come from the development cohort, i.e., 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 single reader may be sufficient.

Testing Procedure. The testing procedure should be designed to demonstrate promising technological 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.

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 re-emphasize 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 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)."

M.2 Technical task-specific evaluation

Objective

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, since 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.

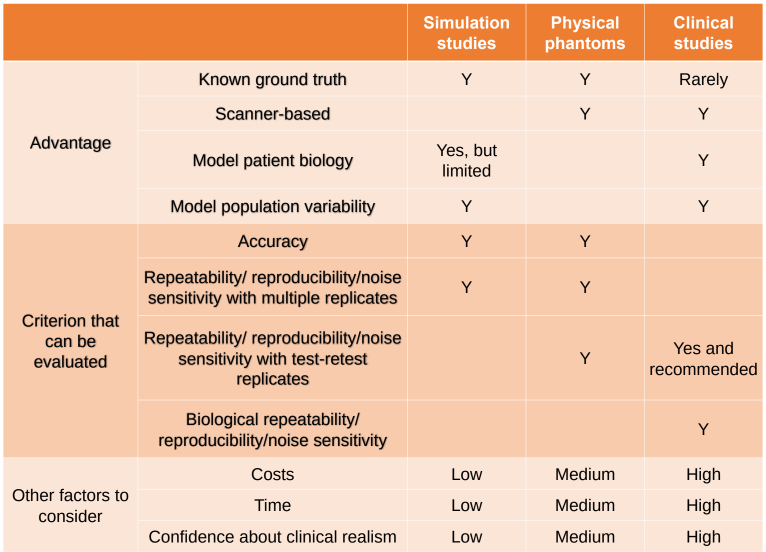
For example, evaluating the accuracy of an AI-based segmentation method to measure metabolic tumor volume (MTV) would be a technical efficacy study. This study would not assess whether more accurate MTV measurement led to any change in clinical outcome.

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 context of PET is provided in Jha et al (*10*).

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 is used to evaluate the technical performance of an AI algorithm. For example, repeatability of an AI algorithm measuring MTV in test-retest PET scans.



		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
	Biological repeatability/ reproducibility/noise sensitivity			Y
Other factors to consider	Costs	Low	Medium	High
	Time	Low	Medium	High
	Confidence about clinical realism	Low	Medium	High

Table 1: Technical evaluation: Comparison of different study types, associated trade-offs, and criterion that can be evaluated with the study type

The tradeoffs with these three study types are listed in Table 1. Each study type can be single or multi-scanner/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.
- Multi-center/multi-scanner 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, multi-center studies are performed to improve patient accrual in trials and therefore the same in- 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 other tools (10) can be used. Different system configurations, including those replicating multi-center 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-phantoms 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 post-reconstruction algorithms should ideally be evaluated using psychophysical studies by expert readers. Methods such as two alternative forced choice tests and ratings-scale approaches could be used. When human-observer studies are infeasible, validated numerical 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. Numerical 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, multi-reader 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 and/or observers should be reported.

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 segmentation algorithm yielded MTV values with a normalized bias of X% (provide 95% confidence intervals) as evaluated using physical-phantom studies with an anthropomorphic thoracic phantom conducted on a single scanner in a single center."

M.3 Clinical evaluation

Objective

Evaluate the impact of the AI algorithm on making clinical decisions, including diagnostic, prognostic, predictive and therapeutic decisions for primary end points such as improved accuracy or precision in measuring clinical outcome. While technical evaluation was geared towards 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.

Study Design

Study Type. Following study types can be used:

- Retrospective study: A retrospective study employs 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 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 non-randomized trials and quasi-experiments 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 post-deployment 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 to a prospective interventional study. Moreover, the real-world data can not only be leveraged to improve performance of the initially cleared AI device but also evaluate new clinical applications that require the same/similar data as 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 prior to retrieving/analyzing the real-world data (*46,47*), with special attention paid to negate bias (*48*).

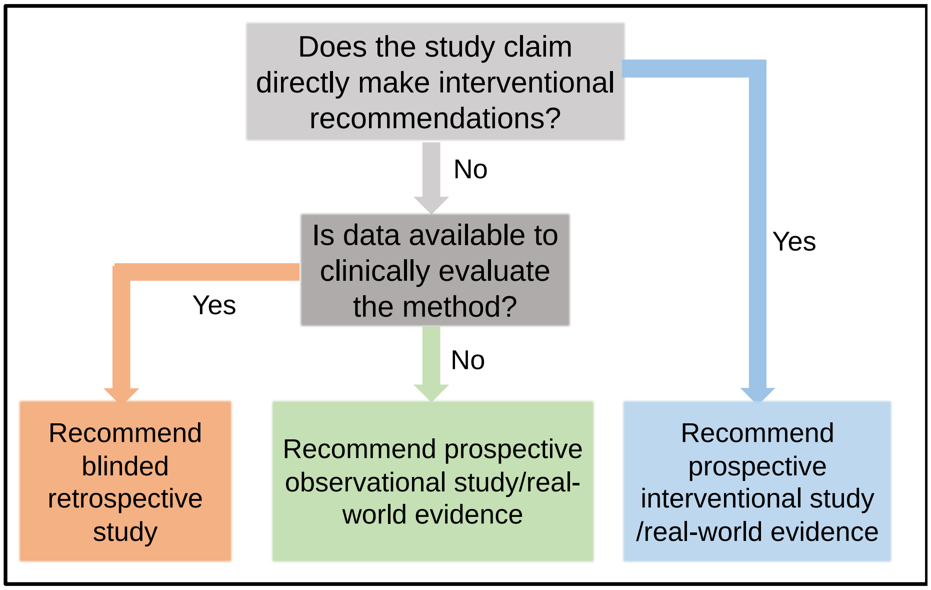

Does the study claim directly make interventional recommendations?
No
Yes
Is data available to clinically evaluate the method?
Yes
No
Recommend blinded retrospective study
Recommend prospective observational study/real-world evidence
Recommend prospective interventional study /real-world evidence

Fig. 5: Flowchart to determine the clinical evaluation strategy

Choosing the study type: This is a multi-factorial 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) vs. those that predict the future (e.g., prognosis of therapy outcome, disease progression, overall survival). There are close links between these two categories, and the line between them will likely be increasingly blurred in the era of AI: e.g., 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, leave-one-site repeated hold-out (for example 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 check 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 use of the AI method vs. 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 two AUCs. Inputs required for power analysis to compute sample size may be obtained from POC and technical evaluation studies or separate pilot studies.

Defining 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 standard. 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 non-essential call back tests with AI).
- External standard: For disease diagnosis tasks, when available, an external standard such as invasive findings, e.g., biopsy-pathology or invasive coronary angiography, or some other definitive diagnosis (derived from other means than the images utilized) should 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 with this protocol. In such a case, agreement-based studies can be performed and concordance with this 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.

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/non-blinded, randomized/non-randomized, retrospective/prospective/post-deployment, observational/interventional, number of readers.
- Process to define reference standard and figure of merit 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 to 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 pathological complete response in patients with stage II/III breast cancer, as evaluated using a non-randomized 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 NSCLC. 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.

M.4. Post-deployment evaluation

Objective

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

Evaluation Strategies

Monitoring. Quality and patient safety are critical factors in post-deployment 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 employed 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 post-deployment 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 standardized uptake value recovery, absence of non-uniformities or artifacts, and cold-spot recovery, and other specialized tests depending on the AI algorithm. Also, tests can be conducted to assure 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 non-stationary environments with shifting patient populations and clinical and operational practices changing over time (*9*). Post-deployment 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 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 of 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*).

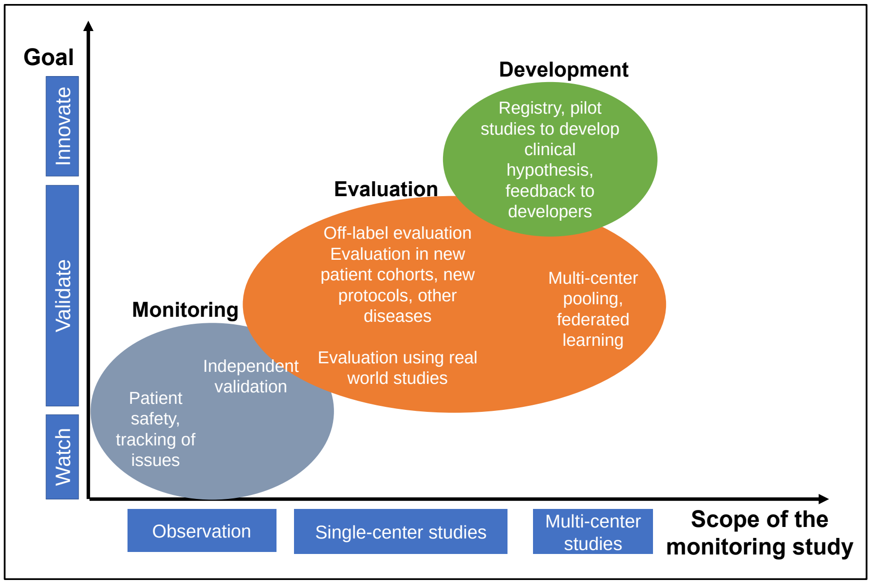

Goal
Innovate
Validate
Watch
Development
Registry, pilot studies to develop clinical hypothesis, feedback to developers
Evaluation
Off-label evaluation
Evaluation in new patient cohorts, new protocols, other diseases
Multi-center pooling, federated learning
Evaluation using real world studies
Monitoring
Independent validation
Patient safety, tracking of issues
Observation
Single-center studies
Multi-center studies
Scope of the monitoring study

Fig. 6: A chart showing the different objectives of post-deployment monitoring, grouped as a function of the scope and goal of the study

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 technological 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, multi-center 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, similar FoMs as 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: Similar FoMs as when performing technical and clinical evaluation.

DISCUSSIONS

The key recommendations from this manuscript 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 towards 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 FDA (*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 four-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 multi-part study.

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 an optimal 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 post-deployment 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.
Post-deployment 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.

Table 2: RELAINCE guidelines

These evaluation studies should be multidisciplinary, and include computational imaging scientists, physicians, physicists, and statisticians right from the study-conception stage. Physicians should be closely involved since 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 semi-supervised approaches. For example, we recommend that for even semi-supervised 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, while 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 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 sub-optimal 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 quality of healthcare.

DISCLOSURES

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 research grant from Siemens Medical Solutions, is a consultant for IBA, and receives royalties from Cedars-Sinai for nuclear cardiology software.

ACKNOWLEDGMENTS

The taskforce 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. *Radiology: Art 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. *J Nucl Med.* 2020;61:575.

4. Leung K, 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 Int 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 18F-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 Clinics.* 2021;16.

11. Barrett HH, Myers KJ. *Foundations of image science.* Vol First: Wiley; 2004.

12. Zhu Y, Yousefirizi F, Liu Z, Klyuzhin I, Rahmim A, Jha A. Comparing clinical evaluation of PET segmentation methods with reference-based metrics and no-gold-standard evaluation technique. *J Nucl Med.* 2021;62:1430-1430.

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 T, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med.* 2021;63.

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 Imag.* 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 Network Open.* 2018;1:e185474-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 Rad Plas 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. *Monte Carlo calculation in nuclear medicine: Applications in diagnostic imaging*; 1998:145-163.

26. Lewellen T, Harrison R, Vannoy S. The SimSET program. *Monte Carlo calculations in nuclear medicine: Applications in diagnostic imaging.* Vol 87; 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 U S A.* 1993;90:9758-9765.

29. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *Journal of Optical Society of America A* 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 90Y 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 Amer A.* 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.

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. 2021:https://arxiv.org/abs/2107.00740.

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.* 2021:1-1.

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 Imag.* 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. *Biochemia medica.* 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.

48. Tarricone R, Boscolo PR, Armeni P. What type of clinical evidence is needed to assess medical devices? *Eur Resp 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 preprint arXiv:181102496.* 2018.

55. Díaz-Rodríguez N, Lomonaco V, Filliat D, Maltoni D. Don't forget, there is more than forgetting: new metrics for Continual Learning. *arXiv preprint arXiv:181013166.* 2018.

56. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:13126211.* 2013.

57. Chaudhry A, Dokania PK, Ajanthan T, Torr PH. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *Proceedings of the European Conference on Computer Vision (ECCV).* 2018:532-547.

58. Sibille L, Seifert R, Avramovic N, et al. 18F-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.

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 Imag.* 2020;7:016502.

61. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology: Art 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.

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.

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.

Nuclear Medicine and Artificial Intelligence - Best Practices for Evaluation (the RELAINCE guidelines)

Supplementary material

A. Example evaluation of AI Application: AI-based transmission-less SPECT reconstruction method

In this supplementary material, we provide an illustration of applying the four-class evaluation framework to evaluate a hypothetical AI-based transmission-less SPECT reconstruction method.

INTRODUCTION

A major imaging-degrading effect in SPECT is the attenuation of gamma-ray photons as they pass through the patient. Attenuation compensation (AC) is considered a pre-requisite for reliable quantification and beneficial for visual interpretation tasks in SPECT (*1*). Typical AC methods require the availability of an attenuation map, often obtained using a transmission scan, such as an X-ray CT scan. However, this has several disadvantages, such as increased radiation dose, higher costs, and possible misalignment between SPECT and CT scans. To address this issue, multiple AI-based transmission-less AC methods for SPECT are being developed. Here we provide a manual to evaluate one such hypothetical AC method using the four-class evaluation framework proposed in the main manuscript. We assume that this hypothetical method has been developed for myocardial perfusion SPECT (MPS). For purposes of illustration, we assume that this method, similar to published approaches (*2,3*), is a deep-learning (DL)-based approach that uses scatter-window projections to estimate the attenuation map. This attenuation map along with the photopeak data are then used to reconstruct the activity map using an ordered subsets expectation maximization (OSEM)-based approach. The manual we provide focuses on the evaluation and not the development of this method. For development, best practices as laid out in Bradshaw et al (*4*) are recommended.

In the discussion below, we will compare our approach with two other AC approaches in SPECT. The first approach uses CT-derived attenuation maps for AC, where the CT can be obtained from a dual-modality SPECT/CT system. This approach is well suited to provide a reference standard when a gold standard is unavailable. The second approach is the Uniform AC method, which uses a uniform attenuation map. The approach is widely used for AC when the attenuation map is unavailable. The Uniform AC method we consider is OSEM-based. In the text below, we denote the deep learning-based AC, CT-based AC, and Uniform AC approaches by DLAC, CTAC, and UniformAC, respectively.

PROOF-OF-CONCEPT EVALUATION

Objective of Evaluation

Demonstrate that the hypothetical DLAC method has promise for further evaluation on clinical tasks.

Study Design

Data collection. For POC evaluation, the evaluator could consider using an existing database of patient images at a medical center on a single scanner. The database should consist of the SPECT projection data in photopeak and scatter windows and the CT scans for these patients, preferably acquired along with the SPECT images. The database should be randomly sampled to define the dataset for this study. This projection data will then be reconstructed using the hypothetical DLAC method to obtain the reconstructed activity images.

Defining reference standard. The reconstructed activity images from the CTAC approach are considered as the reference standard.

Testing procedure. To demonstrate technological innovation, the evaluator should evaluate their method with state-of-the-art and with commonly used methods. The state-of-the-art method would be the CTAC-based approach. The commonly used method would be the UniformAC approach that is OSEM based and assumes a uniform attenuation map. The activity map derived using the DLAC and UniformAC approaches should be compared with the reference standard CTAC-based approach.

Figure of merit. The FoMs to demonstrate technological innovation and promise could include the normalized root mean square error (RMSE), structural similarity index (SSIM), and peak signal-to-noise-ratio (PSNR), along with the corresponding confidence intervals.

Example Claim

A deep learning-based transmission-less SPECT reconstruction method for myocardial perfusion SPECT evaluated on patients acquired on a single scanner from a single center yields SSIM of Y (95% CI) and PSNR of Z dB (95% CI) with the reference standard as CTAC method. The proposed method significantly outperformed the UniformAC method in terms of SSIM and PSNR (*p*-value < 0.05).

TECHNICAL TASK-SPECIFIC EVALUATION

Objective of Evaluation

A major clinical task for which MPS images are acquired is detecting perfusion defects. We describe the procedure to quantify technical efficacy on this detection task.

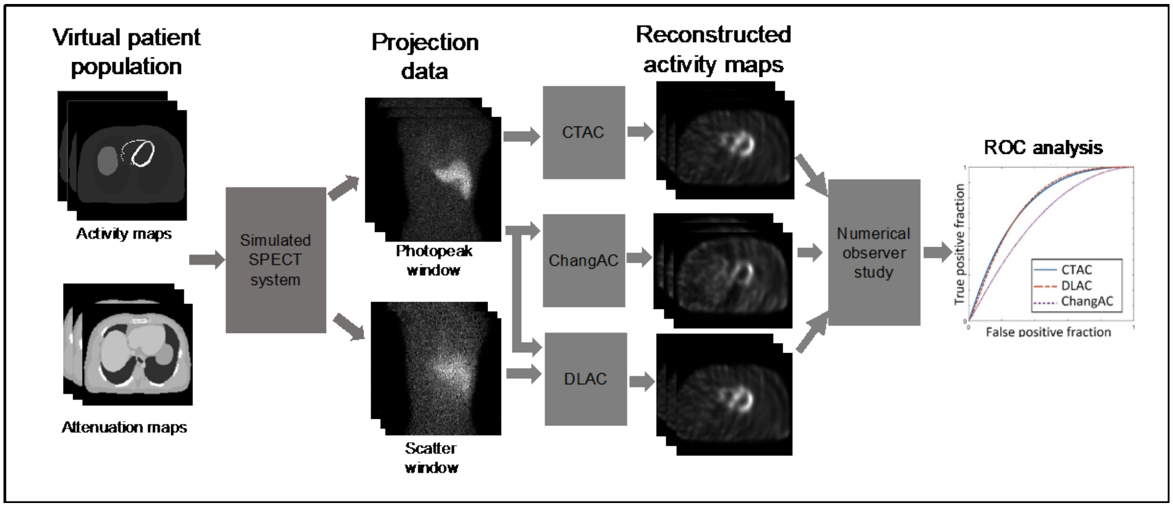
Study Design

A virtual clinical trial provides a rigorous mechanism to conduct this technical evaluation. We describe the study design for a virtual clinical trial-based evaluation:

Data collection. Anthropomorphic phantoms, such as the 3-D extended cardiac and torso phantom, can be used to generate the ground truth patient activity and attenuation maps. The generated patient population should preferably be representative of those seen in clinical practice and have variation in biological properties, including height, weight, and organ shapes and sizes. The patient population should consist of those with and without cardiac defects and prevalence of the defect should preferably be as observed in clinical practice. For the purpose of having a clinical realistic defect variation, defects of different sizes, severities and locations should be simulated. Tracer uptakes should be assigned to various region, according to clinical guided distributions, yielding the simulated digital activity maps. The true attenuation maps can be generated using the 3-D extended cardiac and torso phantom, where the attenuation coefficients are defined at 140 keV, since the tracer used in MPS studies, Tc-99m, emits photons at that energy.

Next, a 3D clinical SPECT system used to acquire MPS images should be accurately simulated. One software to simulate these systems accurately is SIMIND, a photon-tracking-based software (*5*). The acquisition process should simulate clinical protocols. MPS scans are typically conducted with low energy high-resolution collimators and with NaI-based detectors. Further, the SPECT projections are often obtained at 60 angles uniformly spaced over 180 degrees from left posterior oblique to right anterior oblique modeling body-contouring orbits. For the DLAC method, projection data should be obtained in both the photopeak (126-154 KeV) and the scatter window (90-122 KeV). The projection data should then be reconstructed using the DLAC, CTAC and UniformAC methods.

The workflow of virtual clinical trial is shown in Supplemental Figure 1.



Supplemental Figure 1. The workflow of the virtual clinical trial.

Defining a reference standard. Since this is a simulation study, the presence or absence of the defect is known and will thus provide the reference standard.

Process to extract task-specific information. In the evaluation study dataset, the defects vary in activity uptake, shape, and locations, leading to signal variability. Similarly, variation in the shapes and sizes of the other organs, activity uptake through the rest of the body, and variation in patient anatomies leads to background variability. Therefore, this is a signal known statistically/background known statistically (SKS/BKS) task. To avoid bias due to observers, we recommend choosing an optimal observer. One such option could be trained nuclear medicine physicians, but that may make these studies logistically challenging. Another option is numerical observers. One such numerical observer was proposed by Li et al. precisely for this SKS/BKS task (*6*). To use this observer, Li et al cropped the reconstructed activity maps with the centroid of heart at the center of images, and then windowed the intensity values so that the range [0, maximum in the heart] was mapped to the range [0,255]. Then, the testing data was divided into sub-ensembles according to the defect types. The numerical observer that is chosen will yield test statistics. By varying a threshold for these test statistics, the images will be classified into diseased and healthy-patient category. Next, using the knowledge of the ground truth, receiver operating characteristic (ROC) curves can be plotted. This observer study can be conducted with both CTAC and UniformAC approach.

Figures of merit. ROC curves. The area under the ROC curve (AUC), along with the corresponding confidence intervals, should be reported for this technical evaluation study. Delong's test can be used to evaluate if the difference in AUCs using the different methods was statistically significant.

Example Claim

A deep learning-based transmission-less SPECT attenuation compensation (AC) method for myocardial perfusion SPECT was non-inferior to a CT-based AC method on the task of detecting perfusion defects with 80% power and a significance level of 5%. The AUC difference was within a pre-defined margin of 0.1/0.05.

CLINICAL EVALUATION

Objective

Evaluate the efficacy of the hypothetical DLAC method for transmission-less AC in MPS in diagnosing patients with coronary artery disease (CAD).

Study Design

Study type. MPS images are acquired to make diagnostic decisions and not direct therapeutic interventional recommendations. Based on the flowchart in Fig. 5 of the main paper, a blinded retrospective study is chosen for clinical evaluation.

Data collection. The collected data should be from an external cohort. One strategy is to first obtain a database of patients who underwent clinical MPS scans. This institution should be different from the institution that provided the data to train the method. The database should again be representative of patient populations, including patients with different ages, sexes, ethnicities, and BMI. The database should contain projection data in photopeak and scatter windows and the CT scans. The database should then be randomly sampled to define the dataset for the evaluation study. The projection data from this dataset are input to the DLAC approach, yielding the activity maps. These projection data are also used to obtain the activity maps with the CTAC and UniformAC approach.

Defining reference standard. Since we do not know if a patient in this database has CAD or not, we need to define a reference standard. For this purpose, one approach is to use the SPECT images reconstructed with the CTAC approach. These images could be evaluated by a panel of physicians to diagnose if the patient has CAD. The physicians would be provided additional information as required to make this diagnostic decision, such as other clinical-test results or past patient history. Based on the panel consensus, the patients are classified as those with positive and negative CAD diagnosis.

Sample size. A power-analysis is recommended to compute the sample size, where the inputs could be from the proof of concept and the technical efficacy studies.

Reader studies. The evaluation study can be a two alternative forced choice study. In this study, one could have a panel of experienced physicians, who were not involved in the development of the algorithm or defining the reference standard, be presented two images: one from a patient with positive CAD diagnosis and the other from a patient with a negative CAD diagnosis. The physicians would be asked to diagnose which of the two patients has CAD. Additional information as required to make this diagnostic decision, such as other clinical-test results or past patient history would be provided to the physicians. With the reference standard obtained as defined earlier, accuracy for this diagnostic task could be calculated. It can be shown that this accuracy is equal to AUC for this task (*7*).

Figure of merit. One choice for FoM is the AUC for diagnosing CAD, which quantifies the accuracy of diagnosis. The confidence intervals should also be reported for the FoM.

Example Claim

The average AUC of three experienced physicians on the task of diagnosing coronary artery disease by reading myocardial perfusion SPECT images increased from X to Y (increase of ΔAUC (95% confidence intervals)) when these images were reconstructed using a deep learning-based transmission-less AC method as compared to UniformAC method, as evaluated in a blinded retrospective study with clinical patient data collected from two institutions. The reference standard for this study was obtained by three separate readers who read the perfusion SPECT images reconstructed with a CT-based AC approach.

POST-DEPLOYMENT MONITORING

Objective

Evaluate the performance of the DLAC method for an off-label study, namely, AC for quantitative dopamine transporter (DaT) scan SPECT.

Evaluation Strategy

As this is a different clinical application, the algorithm first needs to be trained. For this purpose, best practices as laid out in Bradshaw et al (*4*) are recommended. Here we focus on the evaluation of the algorithm. We will lay out a strategy for technical task-specific evaluation, where the clinical task is to quantify the activity in the putamen and caudate.

Data collection. The dataset used in the off-label evaluation could be from a DaTscan SPECT patient data repository collected on a single scanner from a single center. The patients in this database should be representative of those seen in clinical practice with variations in biological properties, such as genders, ages, ethnicities, and head sizes. The database needs to be randomly sampled to select patients. For the selected patients, the CT images, and projection data both in photopeak (143-175 KeV) and scatter windows (90 – 143 keV) would be selected.

The projection data is then reconstructed using the DLAC, CTAC and UniformAC methods following a similar approach as described in the previous sections but following the clinical protocols for a DaTscan SPECT study.

Defining the reference standard. The reference standard for this quantification task is the uptake in the caudate and putamen region. Since this is a clinical study, the ground-truth uptake values are unavailable. To address this issue, the reference standard can be defined from the images reconstructed using the CTAC approach. To define the reference standard, the caudate and putamen regions need to be segmented from the DaTscan SPECT images. For this purpose, a consensus-based study may be considered where a panel of physicians provide a consensus segmentation for these regions on images obtained with the CTAC approach. The mean activity uptake in the defined left/right caudate and putamen would then define a reference standard.

Process to extract task-specific information. Our goal here is to estimate the uptake in the caudate and putamen region from these images. For this purpose, on the reconstructed images, we could have a panel of physicians, who were not involved in training the method or defining the reference standard, define the boundaries of the caudate and putamen regions. The uptake in these regions will provide the required quantitative values. The same approach could be followed for the images reconstructed with the UniformAC approach.

Figure of merit. Ensemble bias and ensemble mean square error of regional activity uptake obtained by the DLAC/UniformAC method compared with the CTAC method, along with the corresponding confidence intervals.

No-gold-standard evaluation. As mentioned in the main text, another approach to evaluate these methods on the quantitative task of measuring regional uptake is no-gold-standard evaluation. In this evaluation, the average activity in each region obtained by the DLAC, UniformAC, and CTAC methods are calculated. These regional uptake values are then input to the no-gold-standard evaluation technique, which can then rank the different methods on the basis of precision without availability of ground-truth quantitative values.

Claim

The normalized bias of regional activity uptake in the striatal regions obtained with an AI-based transmission-less AC method was X% (95% C.I.) as evaluated in a blinded retrospective study conducted by three readers with data from a repository of patients who underwent DaTscan SPECT on a single scanner in a single center, and where the reference standard was defined as the striatal uptake values computed on the images reconstructed with CT-based AC. Further, the method significantly outperformed the UniformAC method on the quantification task (*p*-value < 0.05).

B. Evaluation of continuous-learning AI-based algorithms

Typically, AI-based clinically available medical devices are locked prior to marketing. However, the performance of these algorithms may degrade when they encounter patient populations, scanners, clinical protocols or other situations different from their training set (*8*). To address this issue, researchers have proposed the continuous-learning (CL) approach (*9*). This approach aims to model the flux or inherent skewness of real-world data to incrementally fine tune model performance. However, CL approaches have to deal with multiple challenges including catastrophic forgetting (whereby, the AI forgets previously learnt information upon learning new information) , skewness in the distribution of the sequentially incoming stream of new data (*9*), and concept drift. Thus, there is an important need for rigorous evaluation of these methods before clinical deployment.

To illustrate an example evaluation strategy, consider an AI-based PET-denoising algorithm that uses the CL approach to account for data drift. The network is deployed at time point t_0. Post-deployment, it is observed that the patient BMIs are more diverse than in the training set. Thus, to account for this change in patient's BMI, the algorithm is retrained at time point t_1. At a later time point t_2, the PET scanner reconstruction algorithms are updated. The PET denoising algorithm is again trained to account for this. Consider an FoM that quantifies performance at each time step on some clinically relevant task. Then we can formulate a 3x3 accuracy matrix (*10*) whose entries, R_{ij}, quantify performance on the test set at time step t_i for the update at time point tj. Using this matrix, we can measure the influence that the retraining has on performance with previous test sets. This performance can be quantified as the average of $R_{1,0}$ - $R_{0,0}$, $R_{2,0}$ – $R_{0,0}$, and $R_{2,1}$ – $R_{1,1}$. This measure, referred as backward transfer, quantifies the forgetting of the AI product through its lifecycle of incremental learning. Analogously, a forward transfer measure can determine the influence that learning a task has on the performance of future tasks (average of the terms $R_{1,0}$, $R_{2,0}$, $R_{2,1}$).

We note that most CL-based deployment insights are in the context of proof-of-concept implementations (*11,12*) and their use for nuclear-medicine requires further research. For CL evaluation, construction of bias-free external test sets and harmonization of data heterogeneity for digital health are needed. Hence, we recommend that a CL-enabled device be evaluated using the framework as discussed in the main paper, with the participation of various stakeholders, who will have to finalize benchmark datasets, FoMs and basic ground rules such as the frequency of updates, test sets, robustness in cyber-security, countermeasures against reverse engineering, traceability of patient data/model parameters and so on at every successive modular update before clinically deploying a CL model. We envision that multi-institutional data repositories such as the Medical Imaging and Data Resource Center, that exhibit optimal standardization, curation and compliance with ethical responsibilities to honor patients' privacy will play a key role in evaluation of CL methods.

Overall, the CL paradigm aims to rectify flaws of the current static AI algorithms in digital healthcare. However, careful evaluation is required to thoroughly validate the use of CL in nuclear medicine.

C. Figures of merit for evaluating performance in proof-of-concept studies

Supplemental Table 1 provides a list of figures of merit (FoMs) for evaluating performance in proof-of-concept studies for different applications of AI.

Supplemental Table 1: A list of FoMs for proof-of-concept evaluation studies

Application	Evaluation figures of merit
Instrumentation	Percent improvement in timing or spatial resolution or sensitivity
Reconstruction and image enhancement	Mean squared error, Structural similarity index, peak signal to noise ratio, Contrast-to-noise ratio
Image registration	Mean squared error, Structural similarity index, Mutual information
Segmentation	Dice scores, Jaccard distance, Hausdroff distance, Fraction of voxels accurately classified

D. Table of figures of merit for evaluating performance on clinical tasks

Supplemental Table 2 provides figures of merit for technical and clinical evaluation. Figures of merit for detection/classification tasks to demonstrate technical efficacy can also be used as figures of merit for clinical evaluation on diagnostic tasks.

Supplemental Table 2: A list of FoMs to evaluate performance on clinical tasks

Type of task	Evaluation criterion	Figure of merit	Description	Range and Target	Notes
2-class classification	Accuracy	Sensitivity/Sensitivity	Sensitivity: Ability to correctly identify positive cases based on a cut-off Specificity: Ability to correctly identify negative cases based on a cut-off	[0; 1] 1	Not influenced by disease prevalence. Requires a priori choice of cut-off. Sensitivity and specificity should be used in conjunction.
		Youden index = sensitivity + specificity -1	Sensitivity + specificity -1	[-1; 1] 1	Not influenced by disease prevalence. Requires a priori choice of cut-off.
		AUC: Area under the ROC curve	Overall classification accuracy, regardless of the cut-off value.	[0; 1] 1	Not influenced by disease prevalence.
		Likelihood ratio for positive test results = sensitivity / (1-specificity)	Likelihood that an image is classified positive in truly positive images compared to negative images	[0; ∞]	Not influenced by disease prevalence. Requires a priori choice of cut-off.
		Likelihood ratio for negative test results = (1-sensitivity) / specificity	Likelihood that an image is classified negative in truly positive images compared to negative images	[0; ∞]	Not influenced by disease prevalence. Requires a priori choice of cut-off.
		F1-score = 2.(precision.recall)/(precision+recall)	A weighted average of precision and recall	[0; 1] 1	F1 ignores the true negatives and is only relevant when the true negatives do not matter
		Balanced accuracy	Average of specificity and sensitivity	[0; 1] 1	Of interest when data is unbalanced; crude measure of accuracy; Requires a priori choice of cut-off
		Matthew's correlation coefficient	MCC= (TP x TN - FP x FN) / squareroot[(TF + FP) x (TP + FN) x (TN + FP) x (TN + FN)]	[-1; 1] 1	Takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used

					even if the classes are of very different sizes. No intuitive interpretation; Requires a priori choice of cut-off
		Positive predictive value (PPV)/Negative predictive value (NPV)	PPV and NPV are probability that cases classified as positive(negative) are truly positive (negative) based on a cut-off, respectively.	[0; 1] 1	Largely influenced by disease prevalence; Requires a priori choice of cut-off
		Precision-recall AUC	Overall classification accuracy, regardless of the cut-off value.	[0; 1] 1	Hypothesis testing methods/software are sparse.
N-class classification	Accuracy	Sensitivity and false positive rate from the N x N confusion matrix	For each class, sensitivity (false positive rate) is the proportion of correctly (incorrectly) classified subjects	[0; 1] 1	Each class has an associated sensitivity and FPR. Requires a priori choice of cut-off. Does not account for types of false classifications.
		Area under the N-dimensional ROC curve	Expansion of the traditional ROC curve to N dimensions	[0; 1] 1	Not influenced by disease prevalence.
		Brier score	Measures accuracy of probabilistic predictions	[0; 1] 0	Can also be applied to 2-class classification
Quantification	Bias	Mean Bias	The mean difference between measured and true value	[-∞; +∞] 0	Unscaled measure of the algorithm's tendency to over- or under-estimate the true value.
		Proportional Bias	Slope of the regression line of true vs measured values	[-∞; +∞] 1	There is proportional bias when slope ≠1 which must be accounted for when measuring change over time.
		Bias profile	Plot of bias over a range of true values		Should be used to evaluate and illustrate when the bias changes over the true value
		Ensemble bias	Average bias over the entire range of true values	[-∞; +∞] 0	Should be used when the bias changes over the true value

	Precision	Standard deviation	Closeness of replicate measurements to each other when repeating the measurements in exactly the same setting	[0; +∞]	Best used when the SD is constant over the range of measurements
		Coefficient of variation	SD divided by the square root of the mean of the measurements	[0; +∞] 0	Best used when the SD is proportional to the magnitude of measurements.
		Precision profile	Plot of standard deviation (or CV) over a range of true values		Should be used when standard deviation (or CV) changes as function of true value
		Ensemble standard deviation	Average standard deviation over the entire range of true values	[0; +∞]	Should be used when standard deviation changes as function of true value
	Reliability	Root Mean Square error	Summary FoM that quantifies both bias and precision	[0; +∞] 0	Informs about bias and variability
	Repeatability Reproducibility	Repeatability Coefficient	Repeatability: Closeness of replicate measurements on the same subject when the same imaging methods were used.	[0; +∞] 0	Describes the smallest difference between two measurements that can be considered a real change with 95% confidence, when there is no change in imaging methods.
		Reproducibility Coefficient	Reproducibility: Closeness of measurements on the same subject when different imaging methods were used (i.e., different scanner, image analysis software, technician, etc).	[0; +∞] 0	Describes the smallest difference between two measurements that can be considered a real change with 95% confidence, when different imaging methods were used.
Quantification	Limits of agreement	Bland Altman analysis	Quantify the agreement between a proposed method and a reference standard		Preferred when the reference standard may be erroneous
Combined detection/ localization	Accuracy	Area under the localization ROC	Accuracy in correctly detecting and locating the lesion	[0; 1] 1	Limited to one lesion per subject
	Accuracy	Area under the FROC curve	Accuracy in correctly detecting and locating lesions	[0; 1] 1	Multiple lesions per subject; summary index difficult to interpret

	Accuracy	Area under the ROI-ROC curve	Accuracy in correctly detecting and locating lesions within mutually exclusive ROIs (e.g. lung lobes, colon segments, breasts)	[0; 1] 1	Multiple lesions per subject; summary index has interpretation similar to traditional ROC area.
	Accuracy	Area under the estimation ROC curve (AUEROC)	Accuracy in correctly detecting and quantifying parameters about the lesion	[0; 1] 1	Generalizes to any joint detection-estimation task
Prediction of Future Events	Probability of occurrence of an event	Survival curve	A plot of the percent of patients that are event-free as a function of time		Can be used for time until any event, such as death, onset of disease, disease re-occurrence.
	Probability of occurrence of an event	Kaplan-Meier estimator	Non-parametric FoM used to estimate the fraction of patients that are event-free at a certain timepoint		Often used to compare survival of two or more cohorts of patients.
	Likelihood of Future event	Prediction risk score	A semi-quantitative risk score that describes the likelihood of a future event taking place based on patient-specific inputs to an algorithm		Binary, ordinal, or continuous value. Often probability based E.g. A score that describes the likelihood of a disease occurring in the future
	Time of future event	Predictive interval	Time interval for which a future event is estimated to occur based on patient-specific inputs to an algorithm		
	Time of future event	Median time of a future event	Median time until future event for typical patient, usually based on longitudinal data from a cohort of patients.		Not patient-specific

REFERENCES

1. Garcia EV. SPECT attenuation correction: an essential tool to realize nuclear cardiology's manifest destiny. *J Nucl Cardiol.* 2007;14:16-24.

2. 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.

3. 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.

4. Bradshaw T, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med.* 2021;63.

5. Ljungberg M, Strand S, King M. The SIMIND Monte Carlo program. *Monte Carlo calculation in nuclear medicine: Applications in diagnostic imaging*; 1998:145-163.

6. 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.

7. Barrett HH, Myers KJ. *Foundations of image science.* Vol First: Wiley; 2004.

8. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med.* 2021;385:283-286.

9. Baweja C, Glocker B, Kamnitsas K. Towards continual learning in medical imaging. *arXiv preprint arXiv:181102496.* 2018.

10. Díaz-Rodríguez N, Lomonaco V, Filliat D, Maltoni D. Don't forget, there is more than forgetting: new metrics for Continual Learning. *arXiv preprint arXiv:181013166.* 2018.

11. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:13126211.* 2013.

12. Chaudhry A, Dokania PK, Ajanthan T, Torr PH. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *Proceedings of the European Conference on Computer Vision (ECCV).* 2018:532-547.