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

Abhinav Jha, Tyler Bradshaw, Irène Buvat, Mathieu Hatt, Prabhat Kc, Chi Liu, Nancy Obuchowski, Babak Saboury, Piotr Slomka, John Sunderland, et al.

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
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 ~ 9000 words

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

Keywords: Artificial intelligence, Clinical task-based evaluation, PET, SPECT, proof of concept, technical efficacy, clinical utility, post-deployment monitoring, generalizability.

Noteworthy:

- We propose a four-class framework to evaluate AI algorithms for nuclear-medicine imaging.
- We provide the RELAINCE (**R**ecommendations for **E**val**u**ation of **AI** for **N**u**c**lear Medicin**e)** guidelines to evaluate promise, technical efficacy, clinical utility, and post-deployment efficacy of AI algorithms.
- We outline key elements that should be specified as the output of an AI-algorithm evaluation study.

Abstract:
An important need exists for strategies for rigorous objective evaluation of artificial intelligence (AI) algorithms for nuclear medicine. To address this need, we propose a four-class framework to evaluate AI algorithms for nuclear medicine. The framework provides a mechanism to evaluate AI algorithms for promise, technical efficacy, clinical utility, and post-deployment efficacy. We provide best practices to evaluate AI algorithms for each of these classes. These best practices are tabulated as a set of RELAINCE (**R**ecommendations for **E**va**l**uation of **AI** for **N**u**c**lear Medicin**e)** guidelines. We recommend that an AI evaluation study should yield a claim and define the key elements of this claim. 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.

A. Introduction

Artificial intelligence (AI)-based algorithms are showing tremendous promise across multiple aspects of nuclear medicine (NM) imaging, including image acquisition, reconstruction, post-processing, diagnostics, prognostics, and clinical decision making. 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, producing tools that are useless or even harmful to patients[1]. The goal of this report is to provide best practices to evaluate AI algorithms in the specific context of NM imaging. We provide these practices in the context of evaluating methods that use artificial neural network (ANN)-based architectures, including deep learning (DL), although many principles are broadly applicable. In the rest of the report, AI-based methods specifically refer to those that use ANNs.

Evaluation has a well-established and essential role in the translation of any imaging technology but is even more critical for AI methods due to their working principles. AI-based methods 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. Thus, the output of these algorithms can be unpredictable. This leads to multiple unique challenges, First, AI algorithms, similar to other imaging methods, may malfunction. For example, AI-based reconstruction may introduce spurious lesions[2] and AI-based lesion segmentation may incorrectly include healthy tissue[3]. Such malfunctioning can adversely impact clinical task performance. Evaluations are thus crucial to assess the algorithm's clinical suitability. A second challenge is that of generalizability. DL-based architectures are highly complicated models with millions of tunable parameters. These methods may perform perfectly in training sets, but not generalize to new data, such as from a different institution[4], different population groups[5, 6] or different scanners[7]. Possible reasons for this include that the algorithm may use data features that correlate with the target outcome only within the training set, or that the training dataset may not be sufficiently representative of a broader patient population. Evaluations are needed to assess the generalizability of these algorithms. A third challenge is that of data drift during clinical deployment. When using AI systems in clinical settings, the input data distribution may drift from the training-data distribution over time due to changes in patient demographics, hardware, image-acquisition and analysis protocols[8]. 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 methods, 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 (SNMMI) put together an Evaluation team in 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 outlining best practices for evaluation of AI methods for nuclear-medicine imaging. This report has been prepared by this team.

In medical imaging, images are acquired for specific clinical tasks. These tasks can be broadly classified into three categories: classification, quantification, or a combination of both. 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, recent studies show that these evaluation strategies may not correlate with clinical task performance[2, 9-12]. One study observed that 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, Yu *et al.* observed that evaluation of an AI-based denoising method for cardiac SPECT using fidelity-based FoMs suggested significantly improved performance compared to without denoising. However, when evaluated on the task of detecting cardiac perfusion defects, the performance of the AI-based denoising method was equivalent, if not worse, to that obtained without applying denoising[9]. Such findings demonstrate that task-agnostic approaches to evaluate AI methods have major limitations. Thus, evaluation strategies that 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 technical factors, such as scanners, acquisition, and analysis protocols. Finally, evaluations should yield quantitative measures of performance to enable clear comparison with standard-of-care approaches and other methods and provide guidance for clinical utility. To account for these factors, we recommend that an evaluation strategy for an AI algorithm 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
- Exact definition of the imaging process (acquisition, reconstruction and analysis protocols)
- The process to extract task-specific information
- Figure of merit to describe task performance, including process to define reference standard

We describe each component in Sec. B. To produce such a claim, we propose an evaluation framework in Sec. C. The framework categorizes the evaluation strategies into four classes: proof-of-concept, technical, clinical and post-deployment evaluation. This framework will guide AI-developers to conduct the evaluation study that provides evidence to support their intended claim. In Sec. C, we also provide the best practices for conducting evaluations for each class. The report finally provides the RELAINCE (**R**ecommendations for **E**va**l**uation of **AI** for **N**u**c**lear Medicin**e)** guidelines, that enlist these best practices.

In this paper, the terms “training”, “validation” and “testing” will be used according to their usual meaning in the AI literature. More specifically, 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. The focus of this paper is on evaluation. The development of AI-based algorithms using the training and validation procedures is described in a companion paper[13].

B. Components of the claim

The claim provides a clear and descriptive characterization of the performance of the AI algorithm. The components of a claim are shown in Fig. 1 and described below.

B.1. Definition of the clinical task: In this paper, the term “task” refers to the clinical goal for which the image was acquired. Broadly, in NM imaging, tasks can be grouped into three categories: classification (which can include lesion detection), quantification, or a combination of both. In the classification task, the patient image is

used to classify the patient into a category. For example, identifying if cancer is present or absent, or the cancer stage from an oncological PET image. Similarly, using an AI algorithm to predict whether a patient is expected to respond to therapy or not would be a classification task. In a quantification task, the patient image is used to quantify some parameter, for example, quantifying standardized uptake value (SUV) in an oncological PET image.

B.2: Patient population for whom the task is defined: The performance of an imaging method can be affected by the physical and statistical properties of the imaged patient population. Results for one population may not necessarily translate to others[4, 6]. Thus, the patient population should be clearly 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. Providing these elements in the claim will inform the generalizability of the method.

B.3. Definition of imaging process: The imaging system, acquisition protocol, and reconstruction and analysis parameters may affect task performance. For example, an AI algorithm evaluated for a high-resolution PET system may not apply to systems with lower resolution, since the method may rely on high-frequency features captured by the high-resolution system[7]. Depending on the method, 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 European Association of Nuclear Medicine Research Ltd (EARL) standards. For example, an AI-based denoising approach for ordered subsets expectation maximization (OSEM)-based reconstructed images may not apply to images reconstructed using filtered-backprojection or even for a different number of OSEM iterations since noise properties change with the number of iterations. Thus, depending on the application, these parameters should be specified in the claim. Further, if the algorithm was evaluated across multiple scanners, or with multiple protocols, that should be specified. This would strengthen confidence in generalizability of the algorithm.

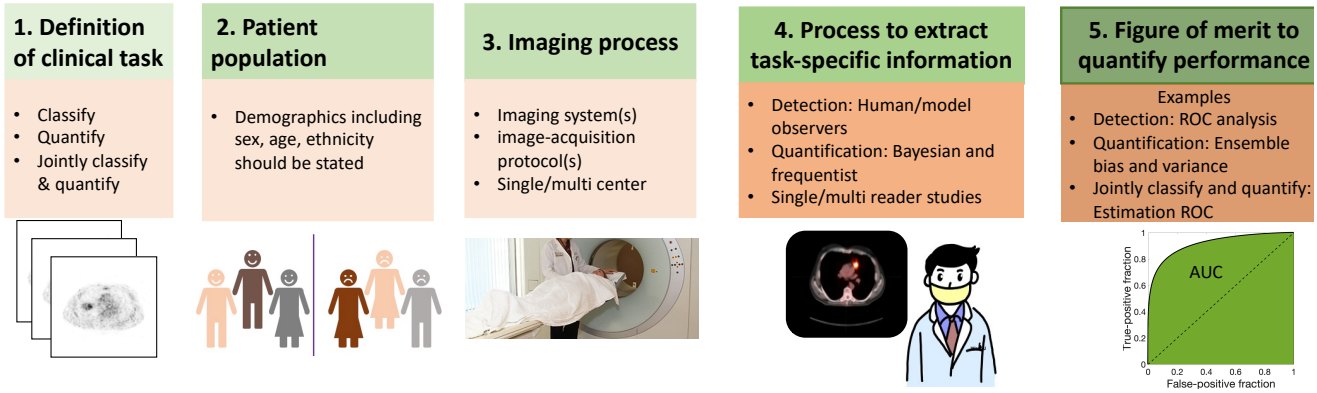

Fig. 1: The components of a claim

B.4. The process to extract task-specific information: Task-based evaluation of an AI imaging algorithm requires a strategy that extracts 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, the competency of the observer (multiple trained radiologists/one trained radiologist/resident/untrained reader) will impact task performance. Further, the choice of the strategy may provide more confidence about 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.

B.5. Figure of merit (FoM) to evaluate task performance: FoMs quantitatively describe a method's performance on the clinical task, enabling comparison of different methods, comparison to standard of care, and helping define quantitative metrics of success. FoMs should be accompanied by confidence intervals

(CIs), which provide a measure of uncertainty in the performance. To obtain the FoM, a reference standard is needed. The process to define the reference standard should be stated.

By providing all the components of a claim, developers will describe the generalizability of the method. Figure 2 presents a schematic that shows how different levels of generalizability can be established. Some key points from this figure are as 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.

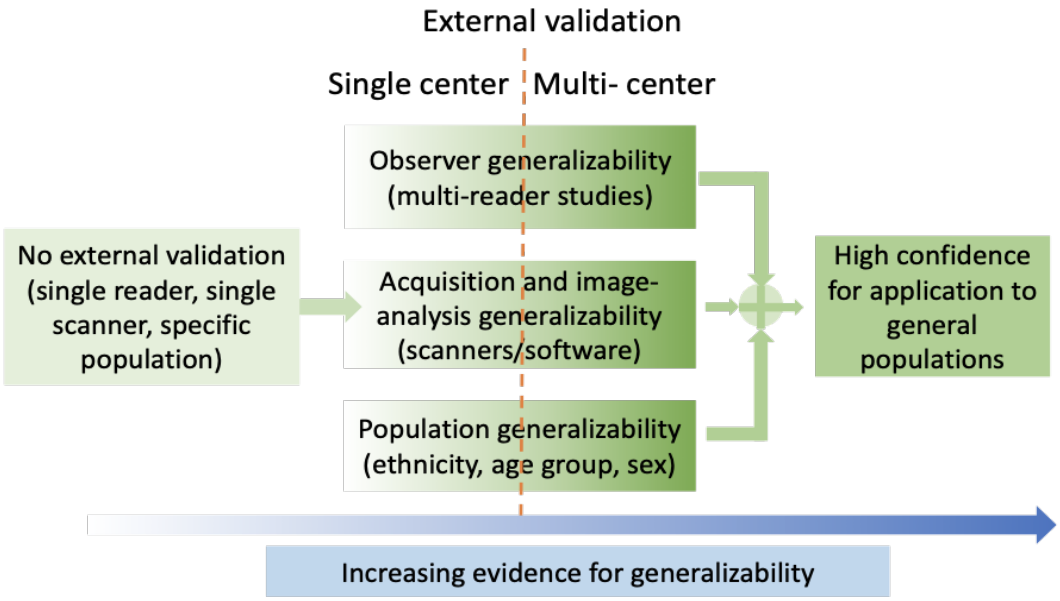

Fig. 2: increasing levels of rigor of evaluation, and how they in turn provide increased confidence in the generalizability

- 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).

C. 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 Sec. C.1-C.4. 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[14] 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 metrics. Suitable for method-development studies. Provides promise for further clinical task-specific evaluation.
  Example: Evaluating the AI PET reconstruction algorithm using root mean square error.
- **Class 2: Technical evaluation**: Quantifies technical performance of an algorithm on a clinical task using measures such as accuracy, repeatability, and reproducibility.

Example: Evaluating performance on the task of lesion detection with the AI low-dose PET reconstructed images using a realistic simulation study.

- **Class 3: Clinical evaluation**: Quantifies the efficacy of the algorithm 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 PET 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 utility of the method in populations and diseases beyond the original claim. It could also include the use of the algorithm with improved imaging cameras and reconstructions which were not used in the original training.
  Example: Evaluating whether the AI PET reconstruction algorithm remains effective over time.

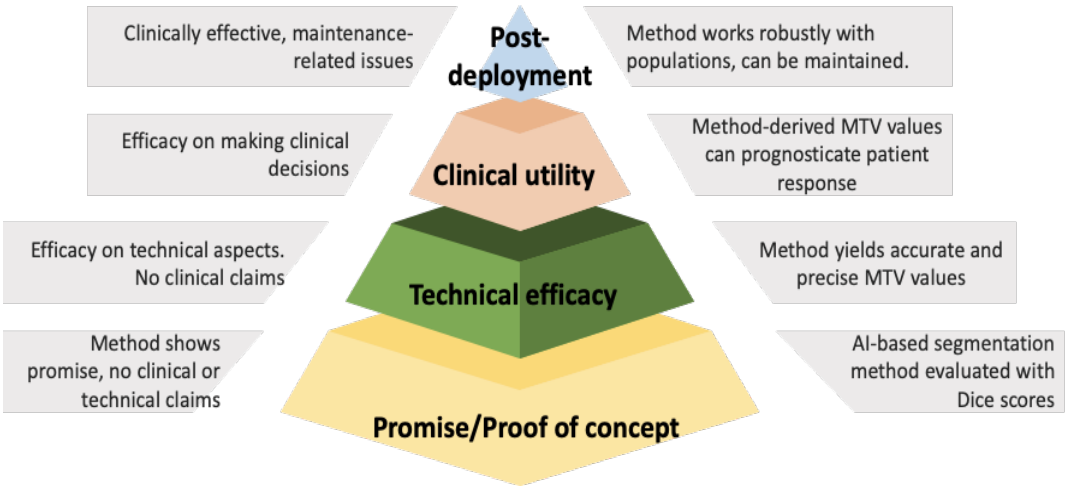

Fig. 3: Framework for evaluation of AI-based algorithms. The left of the pyramid provides a brief description of the phase, and the right provides an example of evaluating an AI-based segmentation algorithm on the task of evaluating metabolic tumor volume (MTV) using this framework.

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 including sample-size considerations,

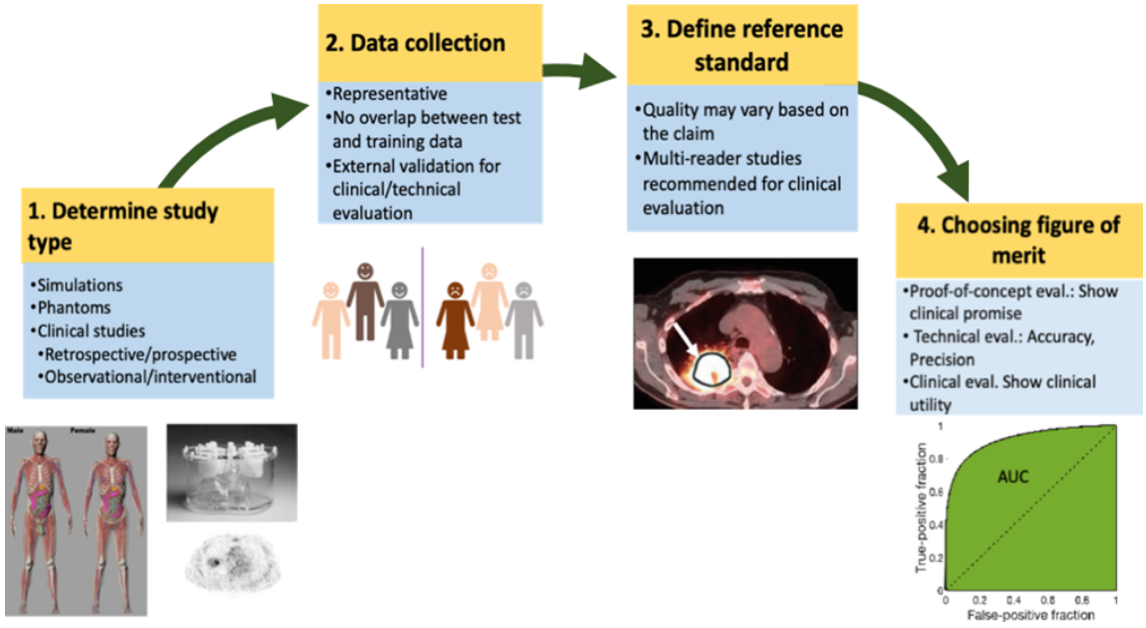

Fig. 4: Elements of study design for each class of evaluation

defining a reference standard, and choosing FoMs (Fig. 4), and finally a generic structure for the claim.

C.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 further 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 method may be suitable for multiple clinical tasks. For example, a segmentation algorithm may be applicable to radiation-therapy planning, estimating volumetric (e.g., metabolic tumor volume (MTV)) or radiomic features, or monitoring therapy response. Evaluating the algorithm on all these tasks would require multiple studies. Further, the developer may not have the necessary resources (such as a large, representative dataset) 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[13].

a. **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, it is important that there is no overlap between these data. Public databases, such as those available at TCIA[15] and from challenges* can also be used.
b. **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. The curation quality need not be of the highest quality. For example, curations by single reader may be sufficient.
c. **Testing procedure**: The testing procedure should be designed to demonstrate promising technological innovation. The algorithm should thus also be compared against a reference or standard of care, and preferably other state-of-the-art algorithms.
d. **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. For this reason, we recommend evaluation of the algorithms using multiple FoMs. The list of these 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 since this is a POC study, the claim should not be interpreted as an indication of the algorithm's expected performance in a clinical setting.
Example claim: Consider the evaluation of a new segmentation algorithm. The claim could read as follows:

* https://grand-challenge.org/challenges/

“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).”

C.2 Technical efficacy 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.

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 is provided in Jha et al[16].

a. **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. A specific class referred to as virtual clinical trials (VCTs) can be used to obtain population-based inferences[17, 18].
   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 study 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
Evaluation criterion	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	Costs	Low	Medium	High
	Time	Low	Medium	High
	Confidence about clinical realism	Low	Medium	High

Table 1: Comparison of different data types and associated trade-offs and evaluation criteria to evaluate technical efficacy

The tradeoffs with these three study types are listed in Table 1. Each study type can be single or multi-scanner/center studies, depending on the claim:

- **Single-center/single-scanner studies** are typically performed with a specific system, image acquisition and reconstruction protocols. 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 technical 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. These studies can identify AI algorithms that are sensitive to variations in scanner performance and reconstruction protocols. Further, multicenter studies can help assess the need for harmonization of imaging procedures and system performances.

b. Data collection:

- **Realistic simulation studies**: To conduct realistic simulations, multiple digital anthropomorphic phantoms are available[19]. 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[3]. Expert reader-based studies can be used to validate realism of simulations[20].
  Next, to simulate the imaging systems, tools such as GATE[21], SIMIND[22], SimSET[23], PeneloPET[24], and other tools[16] 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 that was used for training/validation of the algorithm.
- **Anthropomorphic physical-phantoms studies**: For clinical relevance, the tracer uptake and acquisition parameters when imaging these phantoms should be similar to 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 such as TCIA may also be used. Selection criteria should be defined.

c. 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 (2-AFC) tests and ratings-scale approaches could be used. When human-observer studies are infeasible, numerical observers, such as the channelized Hotelling observer, could be used[25-27]. This is a better choice than using human observers with limited training, who may yield misleading measures of task performance. AI algorithms for optimizing instrumentation/acquisition can be evaluated directly on projection data, which provides the benefit that the evaluation would be agnostic to the choice of the reconstruction and analysis method[28, 29]. 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[25]. The ideal observer can be computationally challenging to obtain in clinical settings, and to address this, different strategies are being developed[30, 31]. 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[16].
- **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. NM 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[16]. If only using the measurements and not incorporating any prior information on the parameters that are quantified, maximum-likelihood estimation methods are an excellent choice[32]. If using prior information, estimators that yield maximum-a-posteriori[33] and posterior-mean[34] estimates could be used. In several cases, measuring quantitative features directly from projection data may yield more reliable quantification[32] 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[35] and those based on deep learning[36] can also be used.

d. **Defining a reference standard:** For simulation studies, the ground-truth is accurately and precisely 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[37]. 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 even without any ground truth[38-40].

e. **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 (EROC) curve for joint detection/classification tasks. 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 or accreditation standards that need to be adhered to for this claim to hold.
- 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 Sec. C.1, 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.”

C.3 Clinical utility evaluation

Objective: Evaluate the utility of an AI algorithm on making clinical decisions, including diagnostic, prognostic, predictive and therapeutic decisions. While technical evaluation was geared towards quantifying the

performance of a technique in controlled settings, clinical evaluation investigates clinical utility in a practical setting.

Study design:

a. **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 bias in patient selection.
- **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[41]. 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 trial (RCTs), although other designs such as non-randomized trials and quasi-experiments are possible[42]. RCTs 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 make use of real-world data (RWD) from AI algorithms that have received regulatory clearance[43]. Such studies have the potential to provide information on a wider patient population that may be difficult to obtain through the prospective interventional study. Moreover, the RWD can not only be leveraged to improve upon the performance of the initially cleared AI device but also be used to evaluate new AI medical applications that require the same/similar data as the initially cleared AI-module, thus saving time and cost. However, critical to this type of study is that its study design be carefully crafted with a study protocol and analysis plan defined prior to retrieving/analyzing the RWD[43, 44]. Special attention should be paid while designing these studies to negate bias[45].

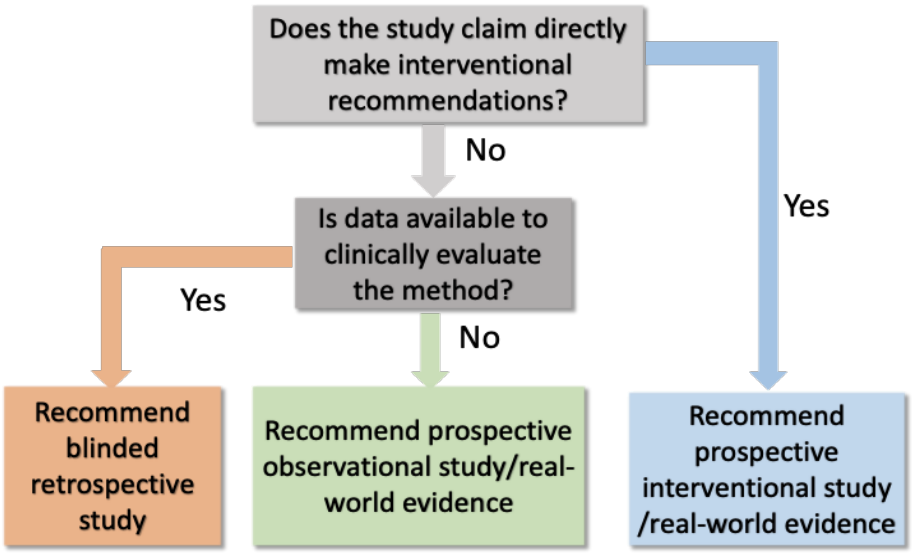

Fig. 5: Flowchart to determine the clinical evaluation strategy

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 any direct interventional recommendations but may alter clinical decision making. The algorithms can be further categorized into those that provide a description about 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.

b. Data collection

An AI algorithm yielding strong performance using data from one institution may perform poorly on data from other institutions[4]. 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[46]. 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 ROC areas. Inputs required for the power-analysis to compute sample size may be obtained from the POC and technical evaluation studies. Pilot studies could also be conducted to estimate sample sizes.

c. 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 as a result 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 in Sec. C.2 should be followed to design these studies. However, note that, unlike technical evaluation, where the goal was restricted to defect detection, 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[47].

d. Figure of merit: These are summarized in supplemental table 2. To evaluate performance on diagnosis tasks, the FoMs of sensitivity, specificity, ROC curves, and AUC can be used. In well-defined populations with known disease prevalence, parameters such at the PPV and NPV may be operationally significant, as well. Since the goal is to demonstrate clinical utility, sensitivity and specificity may be clinically more relevant than ROC analysis. To demonstrate clinical utility in predictive and prognostic tasks, 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 clinical utility.

Example claims:

i. **Retrospective study**: The average AUC of 3 experienced readers on the task of detecting obstructive coronary artery disease from myocardial perfusion imaging (MPI) PET scans improved from X to Y, representing an estimated difference of Δ (95% CI for Δ), when using an AI-based computer aided diagnosis (CAD) 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 images 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 using 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.

C.4. Post-deployment evaluation

Objective: Post-deployment evaluation has multiple objectives including monitoring algorithm performance post clinical deployment, off-label evaluation, and collecting feedback for proactive development (Fig. 6).

Evaluation strategies:

a. **Monitoring**: Critically important in post-deployment monitoring of an AI method is quality and patient safety. 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 a particular AI method. A simple use case is logging the frequency with which an AI algorithm is used 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-method performance and how this impacted patient and disease management. For wider monitoring, developers should also seek feedback from customers, including focus groups, customer complaint and inquiry tracking, and ongoing technical performance benchmarking[48]. This approach may provide additional evidence on algorithm performance and could assist in finding areas of improvements, clinical needs not yet 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 provide a mechanism for post-deployment evaluation, in particular as sanity checks to assess that the AI algorithm was not affected by a maintenance operation such as a software update. These studies could include assessing contrast or SUV recovery, absence of 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 that is generated in non-stationary environments with shifting patient populations and where clinical and operational practices change over time[8]. Post-deployment studies may be needed to identify these dataset shifts and assess if recalibration or retraining of the AI method may be necessary to maintain performance[49] [50]. Monitoring the distribution of various descriptors of the patient population, including the demographics and the prevalence of the disease can provide cues for detection of dataset shifts. In case of changes in these demographics, 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[51]. In supplemental section B, we discuss strategies[52-54] to evaluate continuous-learning-based methods.

b. **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

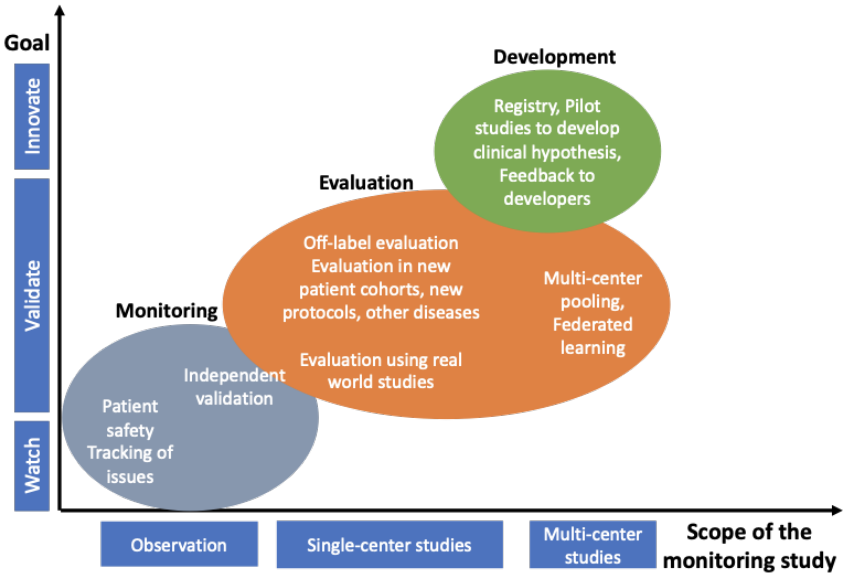

Fig. 6: An eye chart showing the different objectives of post-deployment monitoring, grouped as a function of the scope and goal of the study

design of the algorithm may suggest that an algorithm may exhibit acceptable performance in patient groups 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[55] showed reliable performance in patients with breast cancer[56]. Evaluation of AI algorithms in off-label cohorts can provide evidence of clinical utility beyond the settings initially targeted.

c. **Collecting feedback for proactive development**: Medical products typically have a lifetime longer than a decade. 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 or particular disease.

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 (Sec. C.1) may be considered. However, in case task-based evaluation is required, FoMs as provided in Sec. C.2 may be used.
- Off-label evaluation: Similar FoMs as when evaluating technical efficacy and clinical utility.

D. Discussions and Summary

The key recommendations from this manuscript are summarized in Table 2. These are referred to as the **RELAINCE** (**R**ecommendations for **E**va**l**uation of **AI** for **N**u**c**lear Medicin**e)** guidelines, with the goal of improving the reliance of AI for clinical applications. Unlike other guidelines for the use of AI in radiology[57-59], 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 (IMDRF) and subsequently adopted by the FDA[60]. 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[61].

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. 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. The types of studies required for an algorithm will depend on the claim. A developer may choose to report POC, technical, and clinical evaluation 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 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.
	Ground truth should be high quality and correspond to the task
	Define 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 clinical utility.
	Figure of merit should reflect performance on clinical decision making.
Post-deployment evaluation	Monitor devices and follow reporting guidelines.
	Consider designing 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. In particular, 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[62], including for task-based evaluation of AI algorithms for nuclear medicine[16].

In summary, AI-based technologies 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 methods, and provide trust for clinical application, ultimately leading to improvements in quality of healthcare.

E. 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.

F. Acknowledgements

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

G. References

1. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nature Medicine, 2019. **25**(1): p. 44-56.
2. Yang, J., et al., *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: Artificial Intelligence, 2020. **3**(2): p. e200137.
3. Leung, K., et al., *A physics-guided modular deep-learning based automated framework for tumor segmentation in PET.* Physics in Medicine and Biology, 2020. **65**(24): p. 245032.
4. Zech, J.R., et al., *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.* PLoS Med, 2018. **15**(11): p. e1002683.
5. Gianfrancesco, M.A., et al., *Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data.* JAMA Internal Medicine, 2018. **178**(11): p. 1544-1547.
6. Noor, P., *Can we trust AI not to further embed racial bias and prejudice?* BMJ, 2020. **368**: p. m363.
7. Reuzé, S., et al., *Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners.* Oncotarget, 2017. **8**(26): p. 43169-43179.
8. Finlayson, S.G., et al., *The Clinician and Dataset Shift in Artificial Intelligence.* New England Journal of Medicine, 2021. **385**(3): p. 283-286.
9. Yu, Z., et al., *AI-based methods for nuclear-medicine imaging: Need for objective task-specific evaluation.* Journal of Nuclear Medicine, 2020. **61**(supplement 1): p. 575.
10. Li, K., et al., *Task-based performance evaluation of deep neural network-based image denoising.* Proc. SPIE Med. Imag., 2021. **11599**.
11. Zhu, Y., et al., *Comparing clinical evaluation of PET segmentation methods with reference-based metrics and no-gold-standard evaluation technique.* Journal of Nuclear Medicine, 2021. **62**(supplement 1): p. 1430-1430.
12. KC, P., et al., *Deep neural networks-based denoising models for CT imaging and their efficacy.* Proc. SPIE Med. Imag., 2021. **11595**: p. 105 - 117.
13. Bradshaw, T., et al., *Nuclear Medicine and Artificial Intelligence: Best Practices for Algorithm Development.* Journal of Nuclear Medicine 2021: p. in review.
14. Garcia, E.V., *SPECT attenuation correction: An essential tool to realize nuclear cardiology's manifest destiny.* Journal of Nuclear Cardiology, 2007. **14**(1): p. 16-24.
15. Clark, K., et al., *The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository.* Journal of Digital Imaging, 2013. **26**(6): p. 1045-1057.
16. Jha, A.K., et al., *Objective task-based evaluation of artificial intelligence-based medical imaging methods: Framework, strategies and role of the physician.* PET Clinics, 2021: p. arXiv:2107.04540.
17. Abadi, E., et al., *Virtual clinical trials in medical imaging: a review.* Journal of Medical Imaging, 2020. **7**(4): p. 042805.
18. Yu, Z., et al., *A physics and learning-based transmission-less attenuation compensation method for SPECT.* Proc. SPIE Med. Imag, 2021. **11595**: p. 1159512.

19. Kainz, W., et al., *Advances in Computational Human Phantoms and Their Applications in Biomedical Engineering—A Topical Review.* IEEE Transactions on Radiation and Plasma Medical Sciences, 2019. **3**(1): p. 1-23.
20. Liu, Z., et al., *Observer study-based evaluation of a stochastic and physics-based method to generate oncological PET images.* Proc SPIE Med Imag, 2021. **11599**: p. 1159905.
21. Jan, S., et al., *GATE: a simulation toolkit for PET and SPECT.* Physics in medicine and biology, 2004. **49**(19): p. 4543-4561.
22. Ljungberg, M., S. Strand, and M. King, *The SIMIND Monte Carlo program.* Monte Carlo calculation in nuclear medicine: Applications in diagnostic imaging. Bristol: IOP Publishing, 1998: p. 145-63.
23. Lewellen, T., R. Harrison, and S. Vannoy, *The SimSET program, in Monte Carlo Calculations in Nuclear Medicine: Applications in Diagnostic Imaging.* Institute of Physics Publication, Bristol, UK.
24. España, S., et al., *PeneloPET, a Monte Carlo PET simulation tool based on PENELOPE: features and validation.* Physics in Medicine and Biology, 2009. **54**(6): p. 1723-42.
25. Barrett, H.H., et al., *Model observers for assessment of image quality.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(21): p. 9758-9765.
26. Abbey, C.K. and H.H. Barrett, *Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability.* Journal of Optical Society of America A 2001. **18**(3): p. 473-88.
27. Barrett, H.H. and K.J. Myers, *Foundations of image science*. Vol. First. 2004: Wiley.
28. Gross, K., et al., *Optimizing a multiple-pinhole SPECT system using the ideal observer*. Medical Imaging 2003. Vol. 5034. 2003: SPIE.
29. Rong, X., M. Ghaly, and E.C. Frey, *Optimization of energy window for 90Y bremsstrahlung SPECT imaging for detection tasks using the ideal observer with model-mismatch.* Medical Physics, 2013. **40**(6): p. 062502.
30. Clarkson, E. and F. Shen, *Fisher information and surrogate figures of merit for the task-based assessment of image quality.* Journal of the Optical Society of America. A, Optics, image science, and vision, 2010. **27**(10): p. 2313-2326.
31. Li, X., et al., *Use of Sub-ensembles and Multi-template Observers to Evaluate Detection Task Performance for Data That are Not Multivariate Normal.* IEEE Transactions on Medical Imaging, 2017. **36**(4): p. 917-929.
32. Carson, R.E., *A Maximum Likelihood Method for Region-of-Interest Evaluation in Emission Tomography.* Journal of Computer Assisted Tomography, 1986. **10**(4): p. 654-663.
33. Whitaker, M.K., E. Clarkson, and H.H. Barrett, *Estimating random signal parameters from noisy images with nuisance parameters: linear and scanning-linear methods.* Opt Express, 2008. **16**: p. 8150-8173.
34. Liu, Z., et al., *A Bayesian approach to tissue-fraction estimation for oncological PET segmentation.* Physics in Medicine and Biology, 2021. **66**(12 Special Issue on Early Career Researchers).
35. Tseng, H.-W., J. Fan, and M.A. Kupinski, *Combination of detection and estimation tasks using channelized scanning linear observer for CT imaging systems.* Proc. SPIE Med Imag, 2015. **9416**: p. 94160H.
36. Li, K., et al., *Supervised learning-based ideal observer approximation for joint detection and estimation tasks.* Proc. SPIE Med. Imag, 2021. **11599**: p. 115990F.
37. Miller, D.P., et al., *Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions.* Proc SPIE Med Imag, 2004. **5372**: p. 173-184.
38. Hoppin, J.W., et al., *Objective Comparison of Quantitative Imaging Modalities Without the Use of a Gold Standard.* IEEE Transactions on Medical Imaging, 2002. **21**(5): p. 441-449.
39. Jha, A.K., B. Caffo, and E.C. Frey, *A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods.* Physics in Medicine and Biology, 2016. **61**(7): p. 2780-800.
40. Jha, A.K., et al., *Practical no-gold-standard evaluation framework for quantitative imaging methods: application to lesion segmentation in positron emission tomography.* Journal of Medical Imaging, 2017. **4**(1): p. 011011.
41. Berger, M.L., et al., *Prospective Observational Studies to Assess Comparative Effectiveness: The ISPOR Good Research Practices Task Force Report.* Value in Health, 2012. **15**(2): p. 217-230.
42. Thiese, M.S., *Observational and interventional study design types; an overview.* Biochemia medica, 2014. **24**(2): p. 199-210.

43. Sherman, R.E., et al., *Real-World Evidence — What Is It and What Can It Tell Us?* New England Journal of Medicine, 2016. **375**(23): p. 2293-2297.
44. US Food Drug Administration, *Use of real-world evidence to support regulatory decision-making for medical devices*, in *Guidance for Industry and Food and Drug Administration staff*. 2017.
45. Tarricone, R., P.R. Boscolo, and P. Armeni, *What type of clinical evidence is needed to assess medical devices?* European Respiratory Review, 2016. **25**(141): p. 259.
46. Hemingway, H., R.D. Riley, and D.G. Altman, *Ten steps towards improving prognosis research.* BMJ, 2009. **339**: p. b4184.
47. Shi, L., et al., *Deep learning-based attenuation map generation for myocardial perfusion SPECT.* European Journal Nuclear Medicine Molecular Imaging, 2020. **47**(10): p. 2383-2395.
48. Larson, D.B., et al., *Regulatory Frameworks for Development and Evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms: Summary and Recommendations.* Journal of the American College of Radiology : JACR, 2021. **18**(3 Pt A): p. 413-424.
49. Davis, S.E., et al., *A nonparametric updating method to correct clinical prediction model drift.* Journal of the American Medical Informatics Association, 2019. **26**(12): p. 1448-1457.
50. Feng, J., *Learning to safely approve updates to machine learning algorithms.* Proc. Conf. on Health, Inference, and Learning, 2021: p. 164-173.
51. Baweja, C., B. Glocker, and K. Kamnitsas, *Towards continual learning in medical imaging.* arXiv preprint arXiv:1811.02496, 2018.
52. Díaz-Rodríguez, N., et al., *Don't forget, there is more than forgetting: new metrics for Continual Learning.* arXiv preprint arXiv:1810.13166, 2018.
53. Goodfellow, I.J., et al., *An empirical investigation of catastrophic forgetting in gradient-based neural networks.* arXiv preprint arXiv:1312.6211, 2013.
54. Chaudhry, A., et al., *Riemannian walk for incremental learning: Understanding forgetting and intransigence.* Proceedings of the European Conference on Computer Vision (ECCV), 2018: p. 532-547.
55. Sibille, L., et al., *18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks.* Radiology, 2020. **294**(2): p. 445-452.
56. Weber, M., et al., *Just another "Clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer.* European Journal of Nuclear Medicine and Molecular Imaging, 2021.
57. Dikici, E., et al., *Integrating AI into radiology workflow: levels of research, production, and feedback maturity.* Journal of Medical Imaging, 2020. **7**(1): p. 016502.
58. Mongan, J., L. Moy, and C.E. Kahn, *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers.* Radiology: Artificial Intelligence, 2020. **2**(2): p. e200029.
59. Omoumi, P., et al., *To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines).* European Radiology, 2021. **31**(6): p. 3786-3796.
60. *Software as a medical device (SaMD): Clinical evaluation*. 2017, Center for Devices and Radiological Health, United States Food and Drug Administration.
61. *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*. 2012, Center for Devices and Radiological Health, USA Food and Drug Administration.
62. Rubin, D.L., *Artificial Intelligence in Imaging: The Radiologist's Role.* Journal of the American College of Radiology, 2019. **16**(9 Pt B): p. 1309-1317.