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

Supplemental 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

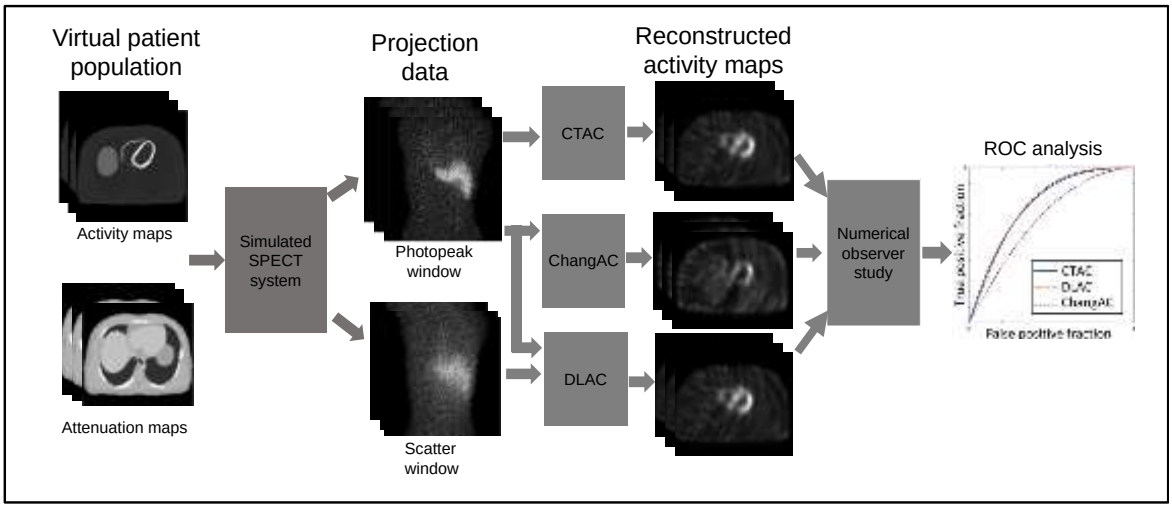
A virtual clinical trial provides a rigorous mechanism to conduct this technical evaluation. We describe the study design for a virtual clinical trial-based evaluation:

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. 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 there is no change in imaging methods.
		Reproducibility Coefficient		[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.